MINIMAX ESTIMATION OF LARGE PRECISION MATRICES WITH BANDABLE CHOLESKY FACTORS

by

Yu Liu

B.S. in Mathematics, Zhejiang University, 2014

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Yu Liu

It was defended on

July 31th 2018

and approved by

Zhao Ren, PhD, Dietrich School of Arts and Sciences, University of Pittsburgh

Satish Iyengar, PhD, Dietrich School of Arts and Sciences, University of Pittsburgh

Sungkyu Jung, PhD, Dietrich School of Arts and Sciences, University of Pittsburgh

Jing Lei, PhD, Dietrich College of Humanities and Social Sciences, Carnegie Mellon University

Dissertation Director: Zhao Ren, PhD, Dietrich School of Arts and Sciences, University of Pittsburgh

MINIMAX ESTIMATION OF LARGE PRECISION MATRICES WITH BANDABLE CHOLESKY FACTORS

Yu Liu, PhD

University of Pittsburgh, 2018

Last decade witnesses significant methodological and theoretical advances in estimating large precision matrices. In particular, there are scientific applications such as longitudinal data, meteorology and spectroscopy in which the ordering of the variables can be interpreted through a bandable structure on the Cholesky factor of the precision matrix. However, the minimax theory has still been largely unknown, as opposed to the well established minimax results over the corresponding bandable covariance matrices. In this thesis, we focus on two commonly used types of parameter spaces, and develop the optimal rates of convergence under both the operator norm and the Frobenius norm. A striking phenomenon is found: two types of parameter spaces are fundamentally different under the operator norm but enjoy the same rate optimality under the Frobenius norm, which is in sharp contrast to the equivalence of corresponding two types of bandable covariance matrices under both norms. This fundamental difference is established by carefully constructing the corresponding minimax lower bounds. Two new estimation procedures are developed: for the operator norm, our optimal procedure is based on a novel local cropping estimator targeting on all principle submatrices of the precision matrix while for the Frobenius norm, our optimal procedure relies on a delicate regression-based block-thresholding rule. Lepski's method is considered to achieve optimal adaptation. We further establish rate optimality in the nonparanormal model, by applying our local cropping procedure to the rank-based estimators. Numerical studies are carried out to confirm our theoretical findings.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 BACKGROUND

Covariance matrix plays a fundamental role in many multivariate statistical problems. They include the principal component analysis, linear and quadratic discriminant analysis, clustering analysis, regression analysis and conditional dependence relationship studies in graphical models. During the last two decades, with the advances of technology, Datasets with high-dimension (the dimension p can be much larger than the sample size n) become common in many applications such as genomics, fMRI data, astrophysics, spectroscopic imaging, risk management, portfolio allocation and numerical weather forecasting [Heyer and Schloerb, 1997, Eisen et al., 1998, Hamill et al., 2001, Ledoit and Wolf, 2003, Schäfer and Strimmer, 2005, Padmanabhan et al., 2016].

1.1.1 Curse of dimensionality

It has been well-known that the sample covariance matrix performs poorly and can yield to invalid conclusions in the high-dimensional settings. For example, see Wachter [1976, 1978], Johnstone [2001], El Karoui [2003], Paul [2007], Johnstone and Lu [2009] for details on the limiting behaviors of the spectra of sample covariance matrices when both n and p increase.

To avoid the curse of dimensionality, certain structural assumptions are almost necessary in order to estimate the covariance matrix or its inverse, the precision matrix, consistently. In this thesis, we consider large precision matrix estimation with bandable Cholesky factor.

1.1.2 Cholesky decomposition

We begin with introducing the bandable Cholesky factor of the precision matrix. Assume that $\mathbf{X} = (X_1, \dots X_p)^T$ is a centered p-variate random vector with covariance matrix Σ. Let $\mathbf{a}_i = (a_{i1}, \dots, a_{i(i-1)})^T$ be the coefficients of the population regression of X_i on its previous variables $\mathbf{X}_{1,i-1} = (X_1, X_2 \dots X_{i-1})^T$. In other words, $\hat{X}_i = \sum_{t=1}^{i-1} a_{it} X_t = \mathbf{X}_{1,i-1}^T \mathbf{a}_i$ is the linear projection of X_i on $\mathbf{X}_{1,i-1}$ in population (Define $\hat{X}_1 = 0$). Set A as the lower triangular matrix with zeros on the diagonal and zero-padded coefficients $(\mathbf{a}_i^T, \mathbf{0})$ arranged in the rows. Denote the residual $\boldsymbol{\epsilon} = \mathbf{X} - \hat{\mathbf{X}} = (I - A)\mathbf{X}$ and $D = \mathsf{Var}(\boldsymbol{\epsilon})$. The regression theory implies the residuals are uncorrelated, and thus the matrix D is diagonal. The modified Cholesky decomposition of Ω is

$$\Omega = \Sigma^{-1} = (I - A)^T D^{-1} (I - A), \tag{1.1}$$

where $I - A$ is the Cholesky factor of Ω. There is a natural order on the variables based on the above Cholesky decomposition. Indeed, the well-known AR(k) model can be characterized by the k-banded Cholesky factor $A \equiv [a_{ij}]_{p\times p}$ of the precision matrix in which $a_{ij} = 0$ if $i - j > k$.

1.1.3 Bandable structure on Cholesky factors

Inspired by the auto-regression model, we consider the bandable structures imposed on the Cholesky factor. More specifically, for $M > 0$, $\eta > 1$ we define the parameter space $\mathcal{P}_\alpha(\eta, M)$ of precision matrices by

$$\begin{aligned} \mathcal{P}_\alpha(\eta, M) = \Big\{\Omega : \quad & \eta^{-1} \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) < \eta, \\ & \max_i \sum_{j<i-k} |a_{ij}| < M k^{-\alpha}, \quad k \in [p]\Big\}. \end{aligned} \tag{1.2}$$

Here, $\lambda_{\max}(\Omega)$, $\lambda_{\min}(\Omega)$ are the maximum and minimum eigenvalues of Ω and the index set $[p] = \{1, 2, \dots, p\}$. We follow the convention that the sum over an empty set of indices is equal to zero when $i - k \le 1$. This parameter space was first proposed in [Bickel and Levina, 2008b]. The parameter α specifies how fast the sequence a_{ij} decays to zero as j goes away from i. The covariance matrix estimation problem has been extensively studied when a similar bandable structure is imposed on the covariance matrix (e.g., [Bickel and Levina, 2008b, Cai et al., 2010]). Unlike the order in these bandable covariance matrices, in which large distance $|i-j|$ implies nearly independence, the order in bandable Cholesky factor encodes nearly conditional independence in the sense that the coefficients a_{ij} is close to zero when $i-j>0$ is large. In the end, besides $\mathcal{P}_\alpha(\eta, M)$, a similar type of classes of parameter spaces with bandable Cholesky factor is considered as well,

$$\begin{aligned}\mathcal{Q}_\alpha(\eta, M) = \Big\{\Omega: \quad & \eta^{-1} \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) < \eta, \\ & |a_{ij}| < M(i-j)^{-\alpha-1}, \quad j \in [i-1]\Big\}.\end{aligned} \tag{1.3}$$

1.1.4 Minimax theory

Minimax theory is a decision rule used in decision theory, which minimizes the possible loss of the worst case scenario. When the goal is to estimate the parameter θ, the performance of an estimator is measured by the loss function $L(\theta, \hat{\theta})$. The risk of the estimator is the expectation of its loss function, $R(\theta, \hat{\theta}) = \mathbb{E}_\theta L(\theta, \hat{\theta})$. The risk function is of help in comparing the performance of different estimators. However, it may not provide a clear answer as to which estimator is better, when neither risk function dominates the other at all values of θ. In order to compare the risk function, one needs an one-number summary to the risk function, and minimax risk is a possible choice. The minimax risk R^* is the infimum of the worst risk over all estimators.

$$R^* = \inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta}).$$

Once the minimax risk R^* has been found, the next step is to find an estimator to achieve that risk. Sometimes the minimax estimator is difficult to find, people settle for the asymptotically minimax estimator, or even the minimax rate estimator.
$\tilde{\theta}$ is a minimax estimator if

$$\sup_{\theta} R(\theta, \tilde{\theta}) = \inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta}).$$

$\tilde{\theta}$ is a asymptotically minimax estimator if

$$\sup_{\theta} R(\theta, \tilde{\theta}) \sim \inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta}), \quad n \to \infty$$

where $a_n \sim b_n$ means $a_n/b_n \to 1$ when n goes to infinity.
$\tilde{\theta}$ is a minimax rate estimator if

$$\sup_{\theta} R(\theta, \tilde{\theta}) \asymp \inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta}), \quad n \to \infty$$

where $a_n \asymp b_n$ means that both a_n/b_n and b_n/a_n are bounded when n goes to infinity.

Note that the minimax estimator is the optimal estimator in the sense of minimax, which performs best in the worst possible case allowed in the problem. Although the minimax estimator is not the perfect estimator, the minimax theory is still meaningful in the theoretical studies. The minimax risk provides the benchmark of the convergence rate by establishing its lower bound which cannot be beaten by any estimator.

Many efforts has been done in the estimation of large matrix in terms of the minimax risks, such as Cai et al. [2010], Cai and Zhou [2012], Cai et al. [2013], Cai et al. [2015], Ren et al. [2015]. The minimax frameworks have been developed in the bandable covariance matrix, the sparse covariance and precision matrix, the Toeplitz covariance matrix, and the sparse spiked covariance matrix under a range of loss functions.

1.1.5 Problem setup

Our goal is to establish the minimax framework of the precision matrix with the bandable Cholesky decomposition factors. The problem is characterized by three ingredients: the parameter space, the loss function and the probability space of the samples.

1.1.5.1 Parameter space Our parameter space consists of the symmetric matrices with the bounded eigenvalues and the bandable Cholesky decomposition factors. Assume that $\Omega = (I-A)^T D^{-1}(I-A)$. For $M > 0$, $\eta > 1$ we define the parameter space $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ of precision matrices by

$$\mathcal{P}_\alpha(\eta, M) = \Big\{\Omega: \quad \eta^{-1} \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) < \eta, \\ \max_i \sum_{j<i-k} |a_{ij}| < M k^{-\alpha}, \quad k \in [p]\Big\}.$$

$$\mathcal{Q}_\alpha(\eta, M) = \Big\{\Omega : \quad \eta^{-1} \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) < \eta, \\ |a_{ij}| < M(i-j)^{-\alpha-1}, \quad j \in [i-1]\Big\}.$$

Here, $\lambda_{\max}(\Omega)$, $\lambda_{\min}(\Omega)$ are the maximum and minimum eigenvalues of Ω and the index set $[p] = \{1, 2, \ldots, p\}$.

It is obvious that $\mathcal{Q}_\alpha(\eta, \alpha M) \subset \mathcal{P}_\alpha(\eta, M)$, note that α has different interpretations in the two parameter spaces. These two parameter spaces are first proposed by Bickel and Levina [2008b], which is applied in modeling the longitudinal and spatial data. Many methodologies has been developed, such as the smoothing method in Wu and Pourahmadi [2003], the penalized likelihood method in Huang et al. [2006], the banded Cholesky decomposition factor in Bickel and Levina [2008b] and the convergence rate of the estimator has been derived. However, the minimax study is not clear in such setting.

1.1.5.2 Loss function To define the loss function we need to define the norm of the matrix first. The induced norm of matrix X is defined as following:

$$\|X\|_q = \sup_{a \neq 0}\{\frac{\|Xa\|_q}{\|a\|_q}\} \tag{1.4}$$

Various norms follow by equation (1.4),

- L_2 norm: $\|X\|_{\text{op}}$ is the largest magnitude of the singular value of matrix X, it is also called operator norm or spectral norm.
- L_1 norm: $\|X\|_1 = \max_{1 \le j \le n} \sum_{i=1}^m |x_{ij}|$, which is the maximum absolute column sum of the matrix.
- L_∞ norm: $\|X\|_\infty = \max_{1 \le i \le m} \sum_{j=1}^n |x_{ij}|$, which is the maximum absolute row sum of the matrix.

Frobenius norm is defined by

$$\|X\|_{\text{F}} = (\sum_{i=1}^p \sum_{j=1}^p a_{ij}^2)^{\frac{1}{2}}, \tag{1.5}$$

in which the matrix is regarded as a long vector.

In the paper, both the operator norm loss ($\|S\|_{\text{op}} = \sup_{\|x\|_2=1} \|Sx\|_2$) and the Frobenius norm loss ($\|S\|_{\text{F}} = (\sum_{i,j} s_{ij}^2)^{\frac{1}{2}}$) are investigated.

$$L_1(\Omega, \tilde{\Omega}) = \|\Omega - \tilde{\Omega}\|_{\text{op}}^2,$$
$$L_2(\Omega, \tilde{\Omega}) = \|\Omega - \tilde{\Omega}\|_{\text{F}}^2.$$

Those two norms are commonly used in practice but with different emphasis. The operator norm reflects the characters of a matrix as an integrity, which is of great importance in the analysis involving the eigen-decomposition, such as principle component analysis, canonical correlation analysis, linear discriminant analysis. While the Frobenius norm focus on the entry-wise performance of the matrix. Note that the minimax risk is totally different under different norms.

1.1.5.3 Probability space Assume that the random variable $\mathbf{X}_{1,p} = (X_1, X_2 \ldots X_p)^T$ with the covariance matrix $\Sigma \equiv [\sigma_{ij}]_{p\times p}$ and the precision matrix $\Omega \equiv [\omega_{ij}]_{p\times p}$, follows the sub-Gaussian distribution with constant $\rho > 0$, that is,

$$\mathbb{P}\{|v'(\mathbf{X} - \mathbb{E}\mathbf{X})| > t\} \leq 2\exp(-t^2\rho/2) \tag{1.6}$$

for all $t > 0$ and all deterministic unit vector $\|v\| = 1$. Without loss of generality, we can always assume that $\mathbb{E}\mathbf{X}_{1,p} = \mathbf{0}$. Let $\mathbf{Z}_{1,p} \in \mathbb{R}^{n\times p}$ is n iid copies of $\mathbf{X}_{1,p}$, as the observation.

Our goal is to estimate the precision matrix Ω over the parameter space $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ defined in equation (1.1.5.1) and (1.1.5.1), under the operator norm and Frobenius norm defined in equation (1.4) and (1.5), by the observation $\mathbf{Z}$.
The following four rates are of crucial concern:

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2,$$
$$\inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2,$$
$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{F}}^2,$$
$$\inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{F}}^2.$$

In this thesis, we provide the corresponding rate-optimal estimation procedures $\tilde{\Omega}$, which require the knowledge of α. The fully data-driven adaptive estimation procedure which is optimal over all the α's is also under consideration. we also investigate the nonparanormal models and establish the rate optimal procedure for estimating the inverse of the correlation matrix.

1.1.6 Challenges

Although several approaches have been developed to estimate the precision matrix with bandable Cholesky factor, the optimality question remains mostly open, partially due to the following two reasons. (i) Intuitively, one would expect the minimax rate of convergence over $\mathcal{P}_\alpha(\eta, M)$ under the operator norm to be the same as that over the class of bandable covariance matrices with the same decay parameter α. Under sub-Gaussian assumptions, Cai et al. [2010] established the optimal rate of convergence $\mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2 \asymp n^{\frac{-2\alpha+1}{2\alpha}} + \frac{\log p}{n}$ uniformly for all bandable covariance matrices $\Sigma = \Omega^{-1} = [\sigma_{ij}]_{p\times p}$ with bounded spectra such that $\max_i \sum_{|j-i|>k} |\sigma_{ij}| < Mk^{-\alpha}$, $k \in [p]$. To establish such a rate of convergence for $\mathcal{P}_\alpha(\eta, M)$, Lee and Lee [2017] provided a lower bound with the matching rate. However, we show a surprising result in this paper that estimation over $\mathcal{P}_\alpha(\eta, M)$ is a much harder task than that over bandable covariance matrices. Therefore, the lower bound in Lee and Lee [2017] is sub-optimal, and all attempts on showing the same rate of convergence $n^{\frac{-2\alpha+1}{2\alpha}} + \frac{\log p}{n}$ intrinsically cannot succeed. (ii) From the methodological aspect, due to the regression interpretation of the Cholesky decomposition (1.1), almost all existing methods rely on an intermediate estimator of A obtained by running regularized regression of each variable against its previous variables $X_i \sim \sum_{j=1}^{i-1} a_{ij}X_j$. For instance, Bickel and Levina [2008b] estimated each row of A by fitting the banded regression model $X_i \sim \sum_{j=\max\{1,i-k\}}^{i-1} a_{ij}X_j$ with some bandwidth k. Wu and Pourahmadi [2003] used an AIC or BIC penalty to pick the best bandwidth k. In addition, Huang et al. [2006] proposed adding a Lasso or Ridge penalty while Levina et al. [2008] proposed using a nested Lasso penalty to the regression. See, for instance, Banerjee and Ghosal [2014], Lee and Lee [2017] for Bayesian approaches following the similar idea. The typical analysis for those estimation procedures in a row-wise fashion is to bound the operator norm by its matrix ℓ_1/ℓ_∞ norm. Although this analysis may provide optimal rates of convergence under the operator norm loss for some sparsity structure (see, i.e., Cai and Zhou [2012], Cai et al. [2016a] for sparse covariance and precision matrices estimation), it might be sub-optimal for the bandable structure as seen in bandable covariance matrix estimation Cai et al. [2010], Bickel and Levina [2008b]. Therefore, in order to obtain rate-optimality over $\mathcal{P}_\alpha(\eta, M)$, a novel analysis or even a new estimation approach is expected.

1.2 MAIN RESULTS

With regard to the above two issues, we provide satisfactory solutions in this paper. We at the first time show that the rate of convergence under the operator norm over $\mathcal{P}_\alpha(\eta, M)$ is intrinsically slower than that over the counterpart class of bandable covariance matrices. This is achieved via a novel minimax lower bound construction. Moreover, in order to obtain a rate-optimal estimator, we propose a novel local cropping estimator which does not rely on any estimator of A, and thus requires a new analysis. Our local cropping approach targets on accurate estimation of principal submatrices of the precision matrix under the operator norm, which results in a tradeoff between one variance term and two bias terms. The name comes after the idea of estimating each principal submatrix of the precision matrix, which is to crop the center k by k submatrix of the inverse of $3k$ by $3k$ sample covariance matrix using their neighbors in two directions of the same size. (During the finalizing process of this paper, we realized that a similar estimator is independently proposed to estimate precision matrices with a different structure [Hu and Negahban, 2017].) Since our procedure does not directly explore the structure on each row of A, the analysis of bias terms is much more involved, requiring a block-wise partition strategy. More details are discussed in Sections 2.1 and 2.2.1.

We further establish another surprising result: the optimal rates of convergence of two spaces, namely $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$, are different under the operator norm. This remarkable distinction is different from the comparison of two similar types of parameter spaces for bandable covariance matrices in Cai et al. [2010] and bandable Toeplitz covariance matrices in Cai et al. [2013]. The contrast of minimax results on $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ is summarized in Theorem 1 below. We mainly focus on the high-dimensional setting, assuming that $\log p = O(n)$ and $n = O(p)$. Otherwise, one can always obtain trivial constant minimax rate (i.e., inconsistency) or the minimax rate as the smaller of p/n and the one in Theorem 1.

Theorem 1. *Under normality assumption, the minimax risk of estimating the precision matrix Ω over the parameter space $\mathcal{P}_\alpha(\eta, M)$ with $\alpha > \frac{1}{2}$ given in (1.1.5.1) under the operator norm satisfies*

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2 \asymp n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n}. \tag{1.7}$$

The minimax risk of estimating the precision matrix Ω over the parameter space $\mathcal{Q}_\alpha(\eta, M)$ with $\alpha > 0$ given in (1.1.5.1) under the operator norm satisfies

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}. \tag{1.8}$$

Moreover, we also consider the minimax rates of convergence of precision matrix estimation under the Frobenius norm loss over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$. This time, we prove that two types of spaces enjoy the same optimal rate of convergence. Together with the different rates of convergence under the operator norm loss, we demonstrate the intrinsic difference between operator norm and Frobenius norm. The Frobenius norm of a p by p matrix is defined as the ℓ_2 vector norm of all entries. Driven by this fact, our estimation approach is naturally obtained by optimally estimating A and D in (1.1) separately. Due to the decay structure in $\mathcal{P}_\alpha(\eta, M)$, which is defined in terms of nested ℓ_1 norm of each row of A, our estimator is based on regression with a delicate block-thresholding rule. The minimax procedure is motivated by wavelet nonparametric function estimation, although the space $\mathcal{P}_\alpha(\eta, M)$ cannot be exactly described by any Besov ball (Cai [2012], Delyon and Juditsky [1996]). We summarize the optimality result under the Frobenius norm in Theorem 2 below.

Theorem 2. *Under normality assumption, the minimax risk of estimating the precision matrix Ω over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ given in (1.1.5.1) and (1.1.5.1) satisfies*

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2 \asymp \inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2 \asymp n^{-\frac{2\alpha+1}{2\alpha+2}}. \tag{1.9}$$

1.3 LITERATURE REVIEW

During the last decade, various structural assumptions are imposed in the literature of high-dimensional statistics in order to estimate the covariance/precision matrix consistently under various loss functions. While mostly driven by the specific scientific applications, popular structures include *ordered sparsity* (bandable covariance matrices, precision matrices with bandable Cholesky factor), *unordered sparsity* (sparse covariance matrices, sparse precision matrices) and other more complicated ones such as certain combination of sparsity and low-rankness (spike covariance matrices, covariance with tensor product, latent graphical models). Many estimation procedures have been proposed accordingly to estimate high-dimensional covariance/precision matrices via taking advantages of these specific structures. For example, banding (Wu and Pourahmadi [2009], Bickel and Levina [2008b], Xiao and Bunea [2014], Bien et al. [2016]) and tapering methods (Furrer and Bengtsson [2007], Cai et al. [2010]) were developed to estimate bandable covariance matrices or precision matrices with bandable Cholesky factor; thresholding procedures were used in Bickel and Levina [2008a], El Karoui [2008], Cai and Liu [2011] to estimate sparse covariance matrices; penalized likelihood estimation (Huang et al. [2006], Yuan and Lin [2007], d'Aspremont et al. [2008], Banerjee et al. [2008], Rothman et al. [2008], Lam and Fan [2009], Ravikumar et al. [2011]) and penalized regression methods (Meinshausen and Bühlmann [2006], Yuan [2010], Cai et al. [2011], Sun and Zhang [2013], Ren et al. [2015]) are designed for sparse precision matrix estimation.

In addition, loss functions critically determine the intrinsic estimation difficult and corresponding efficient estimation procedures. For matrix estimation, important loss functions include operator norm loss ($\|S\|_{\mathrm{op}} = \sup_{\|x\|_2=1} \|Sx\|_2$ for a matrix $S \equiv [s_{ij}]_{p\times q} \in \mathbb{R}^{p\times q}$), Frobenius norm loss ($\|S\|_{\mathrm{F}} = (\sum_{i,j} s_{ij}^2)^{\frac{1}{2}}$) and its equivalent forms such as Bregman divergence loss. Among all losses, the operator loss is arguably the most important one which is regarded as a truly "two-directional" problem because it cannot be essentially reduced to a problem of estimating a single or multiple vectors.

The fundamental difficulty of various covariance/precision matrices estimation problems have been carefully investigated in terms of the minimax risks under the operator norm loss among other losses, especially for those *ordered and unordered sparsity structures*. Specifically, for unordered structures, Cai and Zhou [2012] considered the problems of optimal estimation of sparse covariance while Cai et al. [2016a] (see Ren et al. [2015] as well) established the optimality results for estimating sparse precision matrices. For ordered structures, Cai et al. [2010] established the optimal rates of convergence over two types of bandable covariance matrices. In addition, with an extra Toeplitz structure, Cai et al. [2013] studied optimal estimation of two types of bandable Toeplitz covariance matrices. However, it was still largely unknown about the optimality results on estimating precision matrices with bandable Cholesky factor. See an exposure paper with discussion Cai et al. [2016b] and references therein on minimax results of covariance/precision matrix estimation under some other losses. In this paper, we provide a solution to this open problem by establishing the optimal rates of convergence over two types of precision matrices with bandable Cholesky factor. Thus, this paper completes the minimaxity results of all four sparsity structures commonly considered in literature.

1.4 OVERVIEW OF THE DISSERTATION

The rest of the paper is organized as follows. Firstly, in Chapter 2, we focus on the estimation of precision matrices under the operator norm. We propose our estimation procedures for precision matrix estimation in Section 2.1. The local cropping estimators are designed for estimating precision matrices under the operator norm. Section 2.2 establishes the optimal rates of convergence under the operator norm for two commonly used types of parameter spaces $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$. A striking difference between two spaces are revealed when considering operator norm loss. The upper bounds are obtained by studying the bias-variance tradeoff of the local cropping estimators in Section 2.2.1. The minimax lower bounds are obtained by testing arguments, which reveal a fundamental and striking difference between two spaces when considering operator norm loss in Section 2.2.2. Section 2.2.3 builds the minimax theory in $\mathcal{Q}_\alpha(\eta, M)$. Section 2.3 considers the adaptive estimation through a variation of Lepski's method under the operator norm. In Section 2.4, we extend the results to nonparanormal models for inverse correlation matrix estimation by applying local cropping procedure to rank-based estimators.

Secondly, we turn to investigate the Frobenius norm in Chapter 3. We develop the regression-based block-thresholding estimators for the Frobenius norm in Section 3.1. Section 3.2 considers rate-optimal estimation under the Frobenius norm via carefully studying the properties of regression-based block-thresholding estimators. The results reveal that the fundamental difficulty of estimation for two parameter spaces are the same when considering Frobenius norm loss.

Finally, Chapter 4 presents the numerical performance of our local cropping procedure to illustrate the difference between two parameter spaces by simulation studies. We also demonstrate the sub-optimality of banding estimators, compared to our optimal procedures. All technical lemmas used in proofs of main results are relegated to the supplement.

1.5 NOTATION

We introduce some basic notations that will be used in the thesis. $\mathbf{1}(\cdot)$ indicates the indicator function while $\mathbf{1}$ indicates the all-ones vector. $\mathsf{sgn}(\cdot)$ indicates the sign function. $\lfloor s \rfloor$ represents the largest integer which is no more than s. $\lceil s \rceil$ represents the smallest integer which is no less than s. Define $a_n \asymp b_n$ if there is a constant $C > 0$ independent of n such that $C^{-1} \leq a_n/b_n \leq C$. For any vector x, $\|x\|_p$ indicates its ℓ_p norm. For any p by q matrix $S = [s_{ij}]_{p\times q} \in \mathbb{R}^{p\times q}$, we use S^T to denote its transpose. The ℓ_p matrix norm is define as $\|S\|_p = \sup_{\|x\|_p=1} \|Sx\|_p$. The ℓ_2 matrix norm is also called the the operator norm or the spectral norm, and denoted as $\|S\|_{\mathrm{op}}$. The Frobenius norm is defined as $\|S\|_{\mathrm{F}} = (\sum_{i,j} s_{ij}^2)^{\frac{1}{2}}$. $\lambda_{\max}(S)$ and $\lambda_{\min}(S)$ are the largest and smallest singular values of S when S is not symmetric. When S is a real symmetric matrix, $\lambda_{\max}(S)$ and $\lambda_{\min}(S)$ denote its largest and smallest eigenvalues. $\mathrm{row}_i(S)$ and $\mathrm{col}_i(S)$ indicate the i-th row and column of matrix S. $a : b$ denotes the index set $\{a, a+1, \ldots, b\}$. $[p]$ is short for the set $1 : p$. For the random vector $\mathbf{X} \in \mathbb{R}^{p\times 1}$ and the data matrix $\mathbf{Z} \in \mathbb{R}^{n\times p}$, $\mathbf{X}_{a:b}$ and $\mathbf{Z}_{a:b}$ indicates the $(a : b)$-th columns of $\mathbf{X}^T$ and $\mathbf{Z}$. For any square matrix S, $\mathrm{diag}(S)$ denotes the diagonal matrix with diagonal entries being those on the main diagonal of S while for any vector v, $\mathrm{diag}(v)$ denotes the diagonal matrix with diagonal entries being v. In the estimation procedure under the operator norm, we use the matrix notation in the form of $S_m^{(k)}$ to facilitate the proof, where S is always a square matrix, m indicates the location information, and (k) indicates that the size of $S_m^{(k)}$ is k. Throughout the paper we denote by C a generic positive constant which may vary from place to place but only depends on α, η, M and possibly some sub-Gaussian distribution constant ρ in (2.7).

2.0 METHODOLOGY AND OPTIMALITY UNDER THE OPERATOR NORM

In this section, we introduce our methodologies over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ for estimating precision matrices under both the operator norm and the Frobenius norm. Assume that $\mathbf{X} = (X_1, \ldots X_p)^T$, a p-variate random vector with mean zero and precision matrix Ω_p. Our estimation procedures are based on its n i.i.d. copies $\mathbf{Z} \in \mathbb{R}^{n \times p}$. We write $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_p)$, where each $\mathbf{Z}_i$ consists of n i.i.d. copies of X_i. Our estimation procedures are different under the operator norm and the Frobenius norm.

2.1 ESTIMATION PROCEDURE

We focus on the estimation problem under the operator norm in this section. As we discussed in the introduction, almost all existing methodologies (Wu and Pourahmadi [2003], Huang et al. [2006], Bickel and Levina [2008b]) directly appeal the Cholesky decomposition of the precision matrix. They first estimate the Cholesky factor A and D by auto-regression and then estimate the precision matrix according to $\Omega = (I - A)^T D^{-1} (I - A)$. The corresponding analysis in the row-wise fashion may not suitable for the operator norm loss. In this paper, we propose a novel local cropping estimator, which focuses on the estimation of Ω directly.

To facilitate the illustration of the estimation procedure, we define two matrix operators. The **cropping operator** is designed to crop the center block out of the matrix. For a p by p matrix $E \equiv [e_{ij}]_{p \times p}$, we define the $k \times k$ matrix $\mathbf{C}_m^k(E) \equiv [c_{ij}]_{k \times k}$, where $1 \leq m \leq p - k + 1$, with

$$c_{ij} = e_{i+m-1, j+m-1}, \text{ when } 1 \leq i, j \leq k. \tag{2.1}$$

The parameter m indicates the location and k indicates the dimension. It is clear that $\mathbf{C}_m^k(E)$ is a principal submatrix of E. The **expanding operator** is designed to put a small matrix onto a large zero matrix. For a k by k matrix, $C \equiv [c_{ij}]_{k\times k}$, define the $p \times p$ matrix $\mathbf{E}_m^p(C) \equiv [e_{ij}]_{p\times p}$, where $1 \le m \le p-k+1$, with

$$e_{ij} = c_{i-m+1,j-m+1}, \text{ when } m \le i, j \le m+k-1, \text{ otherwise } e_{ij} = 0. \tag{2.2}$$

The parameter m indicates the location and p indicates the dimension. Note that for a k by k matrix C, we have $\mathbf{C}_m^k(\mathbf{E}_m^p(C)) = C$. An illustration of two operators is provided in Figure 1.

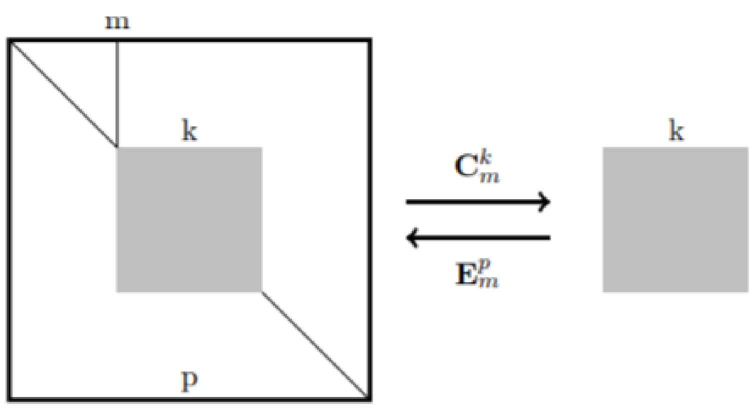

Figure 1: An illustration of the cropping operator and the expanding operator.

In addition, for technical reasons (of obtaining rates of convergence in expectation rather than in probability), we introduce a **projection operator**. For a real square matrix S, let the singular value decomposition of S be $S = U\Lambda V^T$ with $UU^T = I$, $VV^T = I$ and $\Lambda = \text{diag}(\lambda_i)$. Let $\Lambda^* = \text{diag}(\lambda_i^*)$, where $\lambda_i^* = \min\{\max\{\lambda_i, \eta^{-1}\}, \eta\}$, then define

$$\mathbf{P}_\eta(S) = U\Lambda^* V^T. \tag{2.3}$$

For a symmetric matrix S, we modify $\mathbf{P}_\eta(\cdot)$ a little bit and define $\mathbf{P}_\eta(S) = U\Lambda^* U^T$, where $S = U\Lambda U^T$ is its eigen-decomposition. Since all eigenvalues of $\mathbf{P}_\eta(\cdot)$ are in the interval $[\eta^{-1}, \eta]$, $\mathbf{P}_\eta(S)$ is always invertible and positive definite.

We are ready to construct the local cropping estimator $\tilde{\Omega}_k^{\text{op}}$ with bandwidth $k < p$. At a high level, we first propose an estimator of each principal submatrix of size k and $2k$ in Ω using cropping and extending operators. Then we arrange over those local estimators to estimate Ω. Since the core idea of estimating those local estimators in our procedure is to crop the inverse of sample covariance matrix with a relatively larger size, we call $\tilde{\Omega}_k^{\text{op}}$ in (2.6) **the local cropping estimator**.

Specifically, we first define an estimator $\tilde{\Omega}_m^{(k)}$ of the principal submatrix $\mathbf{C}_m^k(\Omega)$ at each location m. To this end, we select the sample covariance matrix with a relative larger size, in this case, $3k$. Let the modified local sample covariance matrix be

$$\tilde{\Sigma}_{m-k}^{(3k)} = \mathbf{P}_\eta\big(\mathbf{C}_{m-k}^{3k}(\frac{1}{n}\mathbf{Z}^T\mathbf{Z})\big). \tag{2.4}$$

Note that the operator $\mathbf{P}_\eta(\cdot)$ guarantees $\tilde{\Sigma}_{m-k}^{(3k)}$ to be invertible. Then we use the center part of its inverse to estimate $\mathbf{C}_m^k(\Omega)$, i.e.,

$$\tilde{\Omega}_m^{(k)} = \mathbf{C}_{k+1}^k\big((\tilde{\Sigma}_{m-k}^{(3k)})^{-1}\big). \tag{2.5}$$

Similarly, we can define local estimators of $\tilde{\Omega}_m^{(2k)}$ via replacing k by $2k$. Arranging over these estimators in the form of weighted sum, we obtain the estimator of Ω, that is,

$$\tilde{\Omega}_k^{\text{op}} = \mathbf{P}_\eta\Big(\frac{1}{k}\big(\sum_{m=2-2k}^{p}\mathbf{E}_m^p(\tilde{\Omega}_m^{(2k)}) - \sum_{m=2-k}^{p}\mathbf{E}_m^p(\tilde{\Omega}_m^{(k)})\big)\Big). \tag{2.6}$$

The operator $\mathbf{E}_m^p(\cdot)$ makes these local estimators in the correct places. The final step (2.6) is motivated by the analysis of optimal bandable covariance matrix estimation procedure proposed in Cai et al. [2010]. Indeed, the optimal tapering estimator in Cai et al. [2010] can be rewritten as a sum of many principal submatrices of the sample covariance matrix in a similar way as (2.6). In contrast, our estimator is not in a form of tapering the sample covariance matrix. However, in the analysis of our local cropping estimator in Section 2.2, the direct target of $\tilde{\Omega}_k^{\text{op}}$ is a certain tapered population precision matrix with bandwidth k. There are natural bias and variance terms involved in the distance of $\tilde{\Omega}_k^{\text{op}}$ and its direct target. Together with the bias of the tapered population precision matrix, our analysis involves two bias terms and one variance term, which critically determine the optimal choice of bandwidth.

In Section 2.2, we show that the local cropping estimator with an optimal choice of bandwidth would achieve the minimax risk under the operator norm over parameter spaces $\mathcal{P}_\alpha(\eta, M)$ in (1.1.5.1) and $\mathcal{Q}_\alpha(\eta, M)$ in (1.1.5.1). However, the optimal choices of bandwidth are fundamentally distinct between $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$. Specifically, we show that the optimal bandwidth over $\mathcal{P}_\alpha(\eta, M)$ is $k \asymp n^{\frac{1}{2\alpha}}$ while that one over $\mathcal{Q}_\alpha(\eta, M)$ is $k \asymp n^{\frac{1}{2\alpha+1}}$.

Remark 1. *Of note, the estimator $\tilde{\Omega}_k^{\mathrm{op}}$ depends on $\mathbf{Z}_{2-4k}, \ldots, \mathbf{Z}_{p+4k-1}$. The index of variable is clear most of the time, while we need to be careful when it is close to the boundary. When the index is beyond the index set $[p]$, we shrink the size of the corresponding block by discarding the data with meaningless indexes.*

2.2 RATE OPTIMALITY UNDER THE OPERATOR NORM

In this section, we establish the optimal rates of convergence for estimating the precision matrix over the parameter spaces $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ given in (1.1.5.1) and (1.1.5.1) under the operator norm. We first derive the risk upper bound of the local cropping estimator in Section 2.2.1 over parameter space $\mathcal{P}_\alpha(\eta, M)$. We provide a matching risk lower bound by applying the Assouad's lemma and the Le Cam's method in Section 2.2.2 over $\mathcal{P}_\alpha(\eta, M)$. The establishment of the rate optimality over the parameter space $\mathcal{Q}_\alpha(\eta, M)$ is similar to the one over $\mathcal{P}_\alpha(\eta, M)$, which is provided in Section 2.2.3.

Throughout this section, we assume that $\mathbf{X} = (X_1, \ldots X_p)^T$ follows certain sub-Gaussian distribution with constant $\rho > 0$, that is,

$$\mathbb{P}\{|v^T(\mathbf{X} - \mathbb{E}\mathbf{X})| > t\} \leq 2\exp(-t^2/(2\rho)), \tag{2.7}$$

for all $t > 0$ and all unit vectors $\|v\|_2 = 1$.

2.2.1 Minimax Upper Bound under the Operator Norm over $\mathcal{P}_\alpha(\eta, M)$

In this section, we develop the following upper bound of our estimation procedure proposed in Section 2.1.

Theorem 3. *When $\lceil n^{\frac{1}{2\alpha}} \rceil \le p$, the local cropping estimator defined in (2.6) of the precision matrix Ω over $\mathcal{P}_\alpha(\eta, M)$ with $\alpha > \frac{1}{2}$ given in (1.1.5.1) satisfies*

$$\sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega}_k^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 \le Ck^{-2\alpha+1} + C\frac{\log p + k}{n}.$$

When $k = \lceil n^{\frac{1}{2\alpha}} \rceil$, we have

$$\sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega}_k^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 \le Cn^{-\frac{2\alpha-1}{2\alpha}} + C\frac{\log p}{n}.$$

The optimal choice of $k \asymp n^{\frac{1}{2\alpha}}$ is due to the bias-variance tradeoff. Combining Theorem 3 with the minimax lower bound derived in Section 2.2.2, we immediately obtain that the local cropping estimator is rate optimal.

Proof. As we discussed in Section 2.1, the direct target of our local cropping estimator is certain tapered population precision matrix with bandwidth k, which can be written as a weighted sum of many principal submatrices of the population precision matrix. We construct this corresponding tapered population precision matrix Ω_k^* as follows. Denote the precision matrix $\Omega \equiv [\omega_{ij}]_{p\times p}$. We define $\Omega_k^* \equiv [\omega_{ij}^*]_{p\times p}$ such that for $i, j \in [p]$,

$$\omega_{ij}^* = m_{ij}\omega_{ij}, \text{ where } m_{ij} = \max\{0, 2 - \frac{1}{k}|i-j|\} - \max\{0, 1 - \frac{1}{k}|i-j|\}. \tag{2.8}$$

The following lemma elucidates the decomposition of this tapered precision matrix Ω_k^*.

Lemma 1. *The Ω_k^* defined in (2.8) can be written as*

$$\Omega_k^* = \frac{1}{k}\Big(\sum_{m=2}^{2k+1}\big(\sum_{j=-1}^{\lfloor p/2k \rfloor}\mathbf{E}^p_{m+2kj}\big(\mathbf{C}^{2k}_{m+2kj}(\Omega)\big)\big) - \sum_{m=2}^{k+1}\big(\sum_{j=-1}^{\lfloor p/k \rfloor}\mathbf{E}^p_{m+kj}\big(\mathbf{C}^{k}_{m+2kj}(\Omega)\big)\big)\Big).$$

The proof of Lemma 1 can be found in [Cai et al., 2010] (refer to the proof of Lemma 1 with covariance matrix therein replaced by the precision matrix), and thus omitted. Define

$$\tilde{\Omega}_k^* = \frac{1}{k}\Big(\sum_{m=2}^{2k+1}\big(\sum_{j=-1}^{\lfloor p/2k\rfloor}\mathbf{E}^p_{m+2kj}(\tilde{\Omega}^{(2k)}_{m+2kj})\big) - \sum_{m=2}^{k+1}\big(\sum_{j=-1}^{\lfloor p/k\rfloor}\mathbf{E}^p_{m+kj}(\tilde{\Omega}^{(k)}_{m+kj})\big)\Big).$$

It is easy to check $\tilde{\Omega}_k^{\mathrm{op}} = \mathbf{P}_\eta(\tilde{\Omega}_k^*)$. Since the eigenvalues of Ω are in the interval $[\eta^{-1}, \eta]$, the operator $\mathbf{P}_\eta(\cdot)$ would not increase the risk much. Indeed, according to (B.1) in Lemma 23, we have

$$\begin{aligned}\mathbb{E}\|\tilde{\Omega}_k^{\mathrm{op}} - \Omega\|^2_{\mathrm{op}} &\le 4\mathbb{E}\|\tilde{\Omega}_k^* - \Omega\|^2_{\mathrm{op}} \\ &\le 8\mathbb{E}\|\tilde{\Omega}_k^* - \Omega_k^*\|^2_{\mathrm{op}} + 8\|\Omega_k^* - \Omega\|^2_{\mathrm{op}}.\end{aligned} \tag{2.9}$$

The following lemma bounds the bias between our direct target Ω_k^* and the population precision matrix.

Lemma 2. *For Ω in the parameter space $\mathcal{P}_\alpha(\eta, M)$ defined in (1.1.5.1) with $\alpha > \frac{1}{2}$, Ω_k^* defined in (2.8), we have*

$$\|\Omega_k^* - \Omega\|^2_{\mathrm{op}} \le Ck^{-2\alpha+1}.$$

Remark 2. *Unlike existing methods, our procedure does not directly utilize the decay structure of Cholesky factor. Consequently, the proof of Lemma 2 is involved and requires a block-wise partition strategy.*

Then we turn to the analysis of $\mathbb{E}\|\tilde{\Omega}_k^* - \Omega_k^*\|^2_{\mathrm{op}}$.

$$\begin{aligned}&\mathbb{E}\|\tilde{\Omega}_k^* - \Omega_k^*\|^2_{\mathrm{op}} \\ \le&2\mathbb{E}\Big(\frac{1}{k}\sum_{m=2}^{2k+1}\|\sum_{j=-1}^{\lfloor p/2k\rfloor}\mathbf{E}^p_{m+2kj}(\tilde{\Omega}^{(2k)}_{m+2kj}) - \sum_{j=-1}^{\lfloor p/2k\rfloor}\mathbf{E}^p_{m+2kj}\big(\mathbf{C}^{2k}_{m+2kj}(\Omega)\big)\|_{\mathrm{op}}\Big)^2 \\ &+ 2\mathbb{E}\Big(\frac{1}{k}\sum_{m=2}^{k+1}\|\sum_{j=-1}^{\lfloor p/k\rfloor}\mathbf{E}^p_{m+kj}(\tilde{\Omega}^{(k)}_{m+kj}) - \sum_{j=-1}^{\lfloor p/k\rfloor}\mathbf{E}^p_{m+kj}\big(\mathbf{C}^{k}_{m+kj}(\Omega)\big)\|_{\mathrm{op}}\Big)^2.\end{aligned} \tag{2.10}$$

These two terms can be bounded in the same way, we only focus on the second term.

$$\begin{aligned}
&\mathbb{E}\big(\frac{1}{k}\sum_{m=2}^{k+1}\|\sum_{j=-1}^{\lfloor p/k\rfloor}\mathbf{E}^p_{m+kj}(\tilde{\Omega}^{(k)}_{m+kj})-\sum_{j=-1}^{\lfloor p/k\rfloor}\mathbf{E}^p_{m+kj}\big(\mathbf{C}^k_{m+kj}(\Omega)\big)\|_{\text{op}}\big)^2\\
\leq&\mathbb{E}\big(\max_m\|\sum_{j=-1}^{\lfloor p/k\rfloor}(\mathbf{E}^p_{m+kj}(\tilde{\Omega}^{(k)}_{m+kj})-\mathbf{E}^p_{m+kj}\big(\mathbf{C}^k_{m+kj}(\Omega)\big))\|^2_{\text{op}}\big)\\
\leq&\mathbb{E}\big(\max_{m,j}\|\tilde{\Omega}^{(k)}_{m+kj}-\mathbf{C}^k_{m+kj}(\Omega)\|^2_{\text{op}}\big)\\
\leq&2\mathbb{E}\big(\max_{m\in[p]}\|\tilde{\Omega}^{(k)}_m-\mathbf{C}^k_{k+1}\big((\mathbf{C}^{3k}_{m-k}(\Omega^{-1}))^{-1}\big)\|^2_{\text{op}}\big)\\
&+2\big(\max_{m\in[p]}\|\mathbf{C}^k_{k+1}\big((\mathbf{C}^{3k}_{m-k}(\Omega^{-1}))^{-1}\big)-\mathbf{C}^k_m(\Omega)\|^2_{\text{op}}\big),
\end{aligned}\tag{2.11}$$

where we further have variance term and bias term of local estimators. For the variance term in (2.11), we further have

$$\begin{aligned}
&\|\tilde{\Omega}^{(k)}_m-\mathbf{C}^k_{k+1}\big((\mathbf{C}^{3k}_{m-k}(\Omega^{-1}))^{-1}\big)\|_{\text{op}}\\
\leq&\|(\tilde{\Sigma}^{(3k)}_{m-k})^{-1}-(\mathbf{C}^{3k}_{m-k}(\Omega^{-1}))^{-1}\|_{\text{op}}\\
\leq&\eta^2\|\tilde{\Sigma}^{(3k)}_{m-k}-\mathbf{C}^{3k}_{m-k}(\Omega^{-1})\|_{\text{op}}\\
\leq&2\eta^2\|\mathbf{C}^{3k}_{m-k}(\frac{1}{n}\mathbf{Z}\mathbf{Z}^T)-\mathbf{C}^{3k}_{m-k}(\Omega^{-1})\|_{\text{op}}.
\end{aligned}\tag{2.12}$$

The last two inequalities hold because of the fact that the eigenvalues of $\tilde{\Sigma}^{(3k)}_{m-k}$ and $\mathbf{C}^{3k}_{m-k}(\Omega^{-1})$ are in the interval $[\eta^{-1},\eta]$, and Lemma 23. The following concentration inequality of sample covariance matrix facilitates our proof.

Lemma 3. *For the observations* $\mathbf{Z}$ *following certain sub-Gaussian distribution with constant* ρ *and precision matrix* Ω*, we have*

$$\mathbb{E}\big(\max_{m\in[p]}\|\mathbf{C}^{3k}_{m-k}(\frac{1}{n}\mathbf{Z}\mathbf{Z}^T)-\mathbf{C}^{3k}_{m-k}(\Omega^{-1})\|^2_{\text{op}}\big)\leq C\frac{\log p+k}{n}.$$

Lemma 3 is an extension of the result in Chapter 2 of [Saulis and Statulevicius, 2012]. Its proof can be found in Lemma 3 of [Cai et al., 2010].

Combining Lemma 3, (2.11) and (2.12), we have

$$\mathbb{E}\big(\max_{m\in[p]}\|\tilde{\Omega}^{(k)}_m-\mathbf{C}^k_{k+1}\big((\mathbf{C}^{3k}_{m-k}(\Omega^{-1}))^{-1}\big)\|^2_{\text{op}}\big)\leq C\frac{\log p+k}{n}.\tag{2.13}$$

We turn to bounding the bias term of local estimator in (2.11).

Lemma 4. *Assume that $\Omega \in \mathcal{P}_\alpha(\eta, M)$ defined in (1.1.5.1) with $\alpha > \frac{1}{2}$. Then we have*

$$\|\mathbf{C}_{k+1}^{k}\big((\mathbf{C}_{m-k}^{3k}(\Omega^{-1}))^{-1}\big) - \mathbf{C}_{m}^{k}(\Omega)\|_{\text{op}}^2 \le Ck^{-2\alpha+1}.$$

Lemma 4, together with (2.13), (2.11) and (2.10), implies that

$$\mathbb{E}\|\tilde{\Omega}_k^* - \Omega_k^*\|_{\text{op}}^2 \le C\frac{\log p + k}{n} + Ck^{-2\alpha+1}. \tag{2.14}$$

Plugging Lemma 2 and (2.14) into (2.9), we finish the proof of Theorem 3. □

2.2.2 Minimax Lower Bound under the Operator Norm over $\mathcal{P}_\alpha(\eta, M)$

Theorem 3 in Section 2.2.1 proves that the local cropping estimator defined in (2.6) attains the convergence rate of $n^{\frac{-2\alpha+1}{2\alpha}} + \frac{\log p}{n}$. In this section, we establish the following matching lower bound, which proves the rate optimality of the local cropping estimator.

Theorem 4. *The minimax risk of estimating the precision matrix Ω over $\mathcal{P}_\alpha(\eta, M)$ defined in (1.1.5.1) under the operator norm with $\alpha \ge \frac{1}{2}$ satisfies*

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \ge \frac{\tau^2}{32}\Big(n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n}\Big), \tag{2.15}$$

where $0 < \tau < \min\{M, \frac{1}{4}\eta^{-1}, \eta^{\frac{1}{2}} - 1\}$.

Remark 3. *Theorems 3 and 4 together show that the minimax risk for estimating the precision matrices over $\mathcal{P}_\alpha(\eta, M)$ stated in (1.7) of Theorem 1. It is worthwhile to notice that there is no consistent estimator over $\mathcal{P}_\alpha(\eta, M)$ under the operator norm, when $\alpha \le \frac{1}{2}$.*

Proof. The lower bound of parameter space $\mathcal{P}_\alpha(\eta, M)$ can be established by the lower bounds over its subsets. We construct two subsets $\mathcal{P}_1$ and $\mathcal{P}_2$ and calculate the lower bound over those two subsets separately. Let τ be a positive constant which is less than $\min\{M, \frac{1}{4}\eta^{-1}, \eta^{\frac{1}{2}} - 1\}$.

First, we construct $\mathcal{P}_1$. Set $k = \min\{\lceil n^{\frac{1}{2\alpha}} \rceil, \frac{p}{2}\}$. Set the index set $\Theta = \{0, 1\}^k$, i.e., for any $\theta \equiv \{\theta_i\}_{1 \le i \le k} \in \Theta$, each θ_i is either 0 or 1. Then we define the $k \times k$ matrix $A_k^*(\theta) \equiv [a_{ij}]_{k \times k}$ with $a_{ij} = \tau n^{-\frac{1}{2}}\theta_i \mathbf{1}\,(j = k)$ and

$$A(\theta) = \begin{bmatrix} 0_{k\times k} & 0_{k\times k} & 0_{k\times(p-2k)} \\ A_k^*(\theta) & 0_{k\times k} & 0_{k\times(p-2k)} \\ 0_{(p-2k)\times k} & 0_{(p-2k)\times k} & 0_{(p-2k)\times(p-2k)} \end{bmatrix}.$$

We then define $\mathcal{P}_1$ as the collection of 2^k matrices indexed by Θ,

$$\mathcal{P}_1 = \left\{ \Omega(\theta) : \Omega(\theta) = (I_p - A(\theta))^T (I_p - A(\theta)), \theta \in \Theta \right\}. \tag{2.16}$$

Next, we construct $\mathcal{P}_2$ as the collection of the diagonal matrices in the following equation,

$$\begin{aligned} \mathcal{P}_2 = \Big\{ & \Omega(m) \equiv [w_{ij}(m)]_{p\times p} : \\ & w_{ij}(m) = \left(\mathbf{1}\,(i=j) + \tau a^{\frac{1}{2}} \mathbf{1}\,(i=j=m) \right)^{-1}, m \in 0 : p \Big\}, \end{aligned} \tag{2.17}$$

where $a = \min\{\frac{\log p}{n}, 1\}$.

Lemma 5. *$\mathcal{P}_1$ and $\mathcal{P}_2$ are subsets of $\mathcal{P}_\alpha(\eta, M)$.*

Note that we assume $\log p = O(n)$ and $n = O(p)$. Without loss of generality, we further assume $\log p < n < p$. For any estimator $\tilde{\Omega}$ based on n i.i.d. observations, we establish the lower bounds over those two subsets in Sections 2.2.2.1 and 2.2.2.2 respectively,

$$\sup_{\mathcal{P}_1} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \geq \frac{\tau^2}{16} n^{-1} \min\{n^{\frac{1}{2\alpha}}, \frac{p}{2}\} \geq \frac{\tau^2}{16} n^{-\frac{2\alpha-1}{2\alpha}}, \tag{2.18}$$

$$\sup_{\mathcal{P}_2} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \geq \frac{\tau^2}{16} n^{-1} \min\{\log p, n\} \geq \frac{\tau^2}{16} \frac{\log p}{n}. \tag{2.19}$$

According to Lemma 5, $(\mathcal{P}_1 \cup \mathcal{P}_2) \subset \mathcal{P}_\alpha(\eta, M)$. Therefore, we obtain

$$\begin{aligned} \sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 &\geq \max\{\sup_{\mathcal{P}_1} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2, \quad \sup_{\mathcal{P}_2} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2\} \\ &\geq \frac{\tau^2}{32} \Big(n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n} \Big), \end{aligned}$$

which completes the proof of Theorem 4. □

We introduce some further notation before establishing (2.18) using Assouad's lemma in Section 2.2.2.1 and (2.19) using Le Cam's method in Section 2.2.2.2. Let $H(\theta, \theta') = \sum_{i=1}^k |\theta_i - \theta_i'|$ be the Hamming distance on $\{0, 1\}^k$, which is the number of different elements between θ and θ'. The total variation affinity $\|P \wedge Q\| = \int p \wedge q \quad d\mu$, where p and q are the density functions of two probability measure P and Q with respect to any common dominating measure μ.

2.2.2.1 Assouad's lemma in proof of (2.18) Assouad's lemma [Assouad, 1983] is a powerful tool to provide the lower bound over distributions indexed by the hypercube $\Theta = \{0,1\}^k$. Let P_θ be the distribution generated from observations indexed by $\Omega(\theta)$. The proof of Lemma 6 can be found in [Yu, 1997], and thus omitted.

Lemma 6 (Assouad)**.** *Let $\tilde{\Omega}$ be an estimator based on observations from a distribution in the collection $\{P_\theta, \theta \in \Theta\}$, where $\Theta = \{0,1\}^k$. Then*

$$\sup_{\theta \in \Theta} 2^2 \mathbb{E}_\theta \|\tilde{\Omega} - \Omega(\theta)\|_2^2 \geq \min_{H(\theta,\theta') \geq 1} \frac{\|\Omega(\theta) - \Omega(\theta')\|_2^2}{H(\theta,\theta')} \frac{k}{2} \min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\|.$$

Applying the Assouad's lemma to the subset $\mathcal{P}_1$, we have the following results.

Lemma 7. *Let P_θ be the joint distribution of n i.i.d. observations from $N(0, \Omega(\theta)^{-1})$, where $\Omega(\theta) \in \mathcal{P}_1$ defined in (2.16). Then*

$$\min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\| \geq 0.5.$$

Lemma 8. *Consider all $\Omega(\theta) \in \mathcal{P}_1$ defined in (2.16). Then*

$$\min_{H(\theta,\theta') \geq 1} \frac{\|\Omega(\theta) - \Omega(\theta')\|_2^2}{H(\theta,\theta')} \geq (\tau n^{-\frac{1}{2}})^2.$$

Lemmas 6, 7 and 8 together imply the desired (2.18), with the choice $k = \lceil n^{\frac{1}{2\alpha}} \rceil$. The proofs of the above lemmas can be found in the supplement.

2.2.2.2 Le Cam's method in proof of (2.19) Le Cam's method can be used to establish the lower bound via testing a single distribution against a convex hull of distributions. Set $r = \inf_{m\in[p]} \|\Omega(0) - \Omega(m)\|_{\text{op}}^2$. Let P_i be the distribution generated from observations indexed by $\Omega(i)$, where $0 \le i \le p$. Define $\bar{P} = \sum_{m=1}^{p} P_m$. The proof of the following lemma can be found in [Yu, 1997], and thus omitted.

Lemma 9 (Le Cam)**.** *Let $\tilde{\Omega}$ be an estimator based on observations from a distribution in the collection $\{P_i, 0 \le i \le p\}$. Then*

$$\sup_{0\le m\le p} \mathbb{E}\|\tilde{\Omega} - \Omega(m)\|_{\text{op}}^2 \ge \frac{1}{2} r \|P_0 \wedge \bar{P}\|.$$

Applying Le Cam's method to $\mathcal{P}_2$, we obtain that $r = \left(\frac{\tau a^{\frac{1}{2}}}{1+\tau a^{\frac{1}{2}}}\right)^2 \ge \frac{1}{4}\tau^2 a$ and the following results.

Lemma 10. *Let P_m be the joint distribution of n i.i.d. observations from $N(0, \Omega(m)^{-1})$, where $\Omega(m) \in \mathcal{P}_2$ defined in (2.17). Then*

$$\|P_0 \wedge \bar{P}\| > \frac{7}{8}.$$

Combining the above results in Lemmas 9 and 10, we obtain the desired (2.19), i.e.,

$$\sup_{0\le m\le p} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \ge \frac{7}{64}\tau^2 a \ge \frac{\tau^2}{16} \min\{\frac{\log p}{n}, 1\}.$$

2.2.3 Rate Optimality under the Operator Norm over $\mathcal{Q}_\alpha(\eta, M)$

2.2.3.1 Minimax upper bound In this section, we establish the optimal rate of convergence over $\mathcal{Q}_\alpha(\eta, M)$ under the operator norm. Our estimation procedure remains the local cropping estimator, except that the bandwidth is $k \asymp n^{\frac{1}{2\alpha+1}}$, which is due to smaller bias terms.

Theorem 5. *When $\lceil n^{\frac{1}{2\alpha+1}} \rceil \le p$, the local cropping estimator defined in (2.6) of the precision matrix Ω over $\mathcal{Q}_\alpha(\eta, M)$ given in (1.1.5.1) satisfies*

$$\sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega}_k^{\text{op}} - \Omega\|_{\text{op}}^2 \le Ck^{-2\alpha} + C\frac{\log p + k}{n}.$$

When $k = \lceil n^{\frac{1}{2\alpha+1}} \rceil$, we have

$$\sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega}_k^{\text{op}} - \Omega\|_{\text{op}}^2 \le Cn^{-\frac{2\alpha}{2\alpha+1}} + C\frac{\log p}{n}.$$

Proof. We employ the same proof strategy as that of Theorem 3. Only two lemmas bounding bias terms need to be replaced. We only emphasize the differences here.

We replace Lemma 2 in the proof by Lemma 11, which bounds the distance of the population precision matrix and its tapered one.

Lemma 11. *For Ω in the parameter space $\mathcal{Q}_\alpha(\eta, M)$ defined in (1.1.5.1), Ω_k^* is defined in (2.8), we have*

$$\|\Omega_k^* - \Omega\|_{\text{op}}^2 \le Ck^{-2\alpha}.$$

In addition, we replace Lemma 4 by Lemma 12, which bounds the bias term of each local estimator.

Lemma 12. *For $\Omega \in \mathcal{Q}_\alpha(\eta, M)$ defined in (1.1.5.1) with $\alpha > 0$, we have*

$$\|\mathbf{C}_{k+1}^k\big((\mathbf{C}_{m-k}^{3k}(\Omega^{-1}))^{-1}\big) - \mathbf{C}_m^k(\Omega)\|_{\text{op}}^2 \le Ck^{-2\alpha}.$$

The remaining part of the proof remains the same, including a similar upper bound for the variance term stated in Lemma 3. Therefore, we complete our proof. $\square$

2.2.3.2 Minimax lower bound

Theorem 6. *The minimax risk for estimating the precision matrix Ω over $\mathcal{Q}_\alpha(\eta, M)$ defined in (1.1.5.1) under the operator norm with $\alpha > 0$ satisfies*

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2 \geq \frac{\tau^2}{32}(n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}). \tag{2.20}$$

Remark 4. *Theorems 5 and 6 together show that the minimax risk for estimating the precision matrices over $\mathcal{Q}_\alpha(\eta, M)$ stated in (1.8) of Theorem 1. In contrast to $\mathcal{P}_\alpha(\eta, M)$, the optimal rate of convergence over $\mathcal{Q}_\alpha(\eta, M)$ is faster. In particular, rate-optimal local cropping estimators are always consistent as long as $\alpha > 0$.*

Proof. To establish the lower bound for $\mathcal{Q}_\alpha(\eta, M)$ in which the decay of a_{ij} is in the entry-wise fashion, we repeat the proof scheme in Section 2.2.2 with a few changes. Let τ be a positive constant which is less than $\min\{M, \frac{1}{4}\eta^{-1}, \eta^{\frac{1}{2}} - 1\}$.

Set $k = \min\{\lceil n^{\frac{1}{2\alpha+1}} \rceil, \frac{p}{2}\}$ and the index set $\Theta = \{0, 1\}^k$, i.e., for any $\theta \in \Theta$, $\theta \equiv \{\theta_i\}_{1 \leq i \leq k}$, each θ_i is either 0 or 1. Define the $k \times k$ matrix $B_k^*(\theta) \equiv [b_{ij}]_{k\times k}$ with $b_{ij} = \tau(nk)^{-\frac{1}{2}}\theta_i$. Define

$$B(\theta) = \begin{bmatrix} 0_{k\times k} & 0_{k\times k} & 0_{k\times(p-2k)} \\ B_k^*(\theta) & 0_{k\times k} & 0_{k\times(p-2k)} \\ 0_{(p-2k)\times k} & 0_{(p-2k)\times k} & 0_{(p-2k)\times(p-2k)} \end{bmatrix}.$$

We construct the collection of 2^k matrices as

$$\mathcal{P}_3 = \left\{\Omega(\theta) : \Omega(\theta) = (I_p - B(\theta))^T (I_p - B(\theta)), \theta \in \Theta\right\}. \tag{2.21}$$

Lemma 13. *$\mathcal{P}_3$ is a subset of $\mathcal{Q}_\alpha(\eta, M)$.*

Let P_θ be the joint distribution of n i.i.d. observations from $N(0, \Omega(\theta)^{-1})$, where $\Omega(\theta) \in \mathcal{P}_3$ defined in (2.21). Parallel to Lemmas 7 and 8 and the lower bound (2.18) for $\mathcal{P}_1$, we establish the following lower bound for $\mathcal{P}_3$.

Lemma 14. *Consider all* $\Omega(\theta) \in \mathcal{P}_3$ *defined in (2.21). Then*

$$\min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\| \geq 0.5, \tag{2.22}$$

$$\min_{H(\theta,\theta')\geq 1} \frac{\|\Omega(\theta) - \Omega(\theta')\|_2^2}{H(\theta,\theta')} \geq (\tau n^{-\frac{1}{2}})^2. \tag{2.23}$$

According to Assouad's lemma, for any estimator $\tilde{\Omega}$ *based on* n *i.i.d. observations, we have*

$$\sup_{\mathcal{P}_3} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \geq \frac{\tau^2}{16} n^{-1} \min\{n^{\frac{1}{2\alpha+1}}, \frac{p}{2}\}. \tag{2.24}$$

It is easy to show $(\mathcal{P}_3 \cup \mathcal{P}_2) \subset \mathcal{Q}_\alpha(\eta, M)$, where $\mathcal{P}_2$ is defined in (2.17). Therefore, combining (2.24) and (2.19), we complete the proof of Theorem 6. □

Remark 5. *The estimation of the covariance matrix* Σ *is of significant importance as well. We propose the estimator of* Σ *by inverting our estimator* $\tilde{\Omega}_k^{\text{op}}$ *given in (2.6). The results and the analysis given in Section 2.2 can be used to establish the minimax optimality of our estimator under the operator norm. According to the inequality* $\|(\tilde{\Omega}_k^{\text{op}})^{-1} - \Sigma\|_{\text{op}} \leq \|(\tilde{\Omega}_k^{\text{op}})^{-1}\|_{\text{op}}\|\tilde{\Omega}_k^{\text{op}} - \Omega\|_{\text{op}}\|\Omega^{-1}\|_{\text{op}}$ *and the fact that both* $\|\tilde{\Omega}_k^{\text{op}})^{-1}\|_{\text{op}}$ *and* $\|\Omega^{-1}\|_{\text{op}}$ *are bounded by* η*, we establish the upper bound of our estimator* $(\tilde{\Omega}_k^{\text{op}})^{-1}$*. Furthermore, considering the analog between the covariance matrix and the precision matrix in the subset* $\mathcal{P}_1$ *and* $\mathcal{P}_2$ *defined in (2.16) and (2.17), the matching lower bound can be proved by a similar argument in Section 2.2.2. Therefore, we have the following rate optimality of estimating the covariance matrix under the operator norm, which can be achieved by estimator* $(\tilde{\Omega}_k^{\text{op}})^{-1}$*,*

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega}^{-1} - \Omega^{-1}\|_{\text{op}}^2 \asymp n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n},$$

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega}^{-1} - \Omega^{-1}\|_{\text{op}}^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}.$$

2.3 ADAPTIVE ESTIMATION

To achieve the minimax rates in Theorem 1 under the operator norm, the local cropping estimator $\tilde{\Omega}_k^{\mathrm{op}}$ requires the knowledge of smoothness parameter α as the optimal choice of bandwidth $k = \lceil n^{\frac{1}{2\alpha}} \rceil$ and $k = \lceil n^{\frac{1}{2\alpha+1}} \rceil$ over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ respectively. In this section, we consider adaptive estimation where the goal is to construct a single procedure which is minimax rate optimal simultaneously over each parameter space $\mathcal{P}_\alpha(\eta, M)$ ($\alpha > 1/2$) and $\mathcal{Q}_\alpha(\eta, M)$ ($\alpha > 0$). Throughout this section, we assume that $\mathbf{X}$ follows certain sub-Gaussian distribution defined in (2.7).

Recall that for each k, the local cropping estimator $\tilde{\Omega}_k^{\mathrm{op}}$ is defined in (2.6). Without the knowledge of α, the bandwidth k needs to be picked in a data-driven fashion. Motivated by the Lepski's methods for nonparametric function estimation problems Lepskii [1992], we select the bandwidth $\hat{k}$ through the following procedure,

$$\hat{k} = \min\{k \in \mathcal{H} : \|\tilde{\Omega}_k^{\mathrm{op}} - \tilde{\Omega}_l^{\mathrm{op}}\|_{\mathrm{op}}^2 \leq C_L \frac{\log p + l}{n}, \text{ for all } l \geq k\}, \tag{2.25}$$

where $\mathcal{H} = \{1, 2, \ldots \lceil \frac{n}{\log p} \rceil\}$ and $C_L > 0$ is a sufficiently large constant. If the set that is minimized over is empty, we use the convention $\hat{k} = \lceil \frac{n}{\log p} \rceil$. The adaptive local cropping estimator $\tilde{\Omega}_{\hat{k}}^{\mathrm{op}}$ enjoys the following theoretical guarantee, and thus is adaptive minimax rate optimal.

Theorem 7. *Assume $\log p = O(n)$, $n = O(p)$. Then the adaptive estimator $\tilde{\Omega}_{\hat{k}}^{\mathrm{op}}$ with $\hat{k}$ defined in (2.25) of the precision matrix Ω over $\mathcal{P}_\alpha(\eta, M)$ with $\alpha > \frac{1}{2}$ satisfies*

$$\sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega}_{\hat{k}}^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 \leq C n^{-\frac{2\alpha-1}{2\alpha}} + C\frac{\log p}{n}.$$

In addition, the adaptive estimator $\tilde{\Omega}_{\hat{k}}^{\mathrm{op}}$ over $\mathcal{Q}_\alpha(\eta, M)$ with $\alpha > 0$ satisfies

$$\sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega}_{\hat{k}}^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 \leq C n^{-\frac{2\alpha}{2\alpha+1}} + C\frac{\log p}{n}.$$

Proof. We only show the upper bound over $\mathcal{P}_\alpha(\eta, M)$ with $\alpha > \frac{1}{2}$. The proof over space $\mathcal{Q}_\alpha(\eta, M)$ with $\alpha > 0$ can be shown similarly and thus omitted.

Set the oracle bandwidth $k^* = \lceil n^{\frac{1}{2\alpha}} \rceil$. For any $\Omega \in \mathcal{P}_\alpha(\eta, M)$, we decompose the risk as follows,

$$\mathbb{E}\|\tilde{\Omega}^{\text{op}}_{\hat{k}} - \Omega\|^2_{\text{op}} \leq 2\mathbb{E}\|\tilde{\Omega}^{\text{op}}_{\hat{k}} - \tilde{\Omega}^{\text{op}}_{k^*}\|^2_{\text{op}} + 2\mathbb{E}\|\tilde{\Omega}^{\text{op}}_{k^*} - \Omega\|^2_{\text{op}}. \tag{2.26}$$

Since k^* is deterministic, we immediately obtain from Theorem 3 that

$$\mathbb{E}\|\tilde{\Omega}^{\text{op}}_{k^*} - \Omega\|^2_{\text{op}} \leq C n^{-\frac{2\alpha-1}{2\alpha}} + C\frac{\log p}{n}, \tag{2.27}$$

which controls the second term of the risk decomposition (2.26).

We turn to bound the first term of (2.26). Due to the definition of $\hat{k}$ and k^*, we have that on the event $\{\hat{k} \leq k^*\}$,

$$\|\tilde{\Omega}^{\text{op}}_{\hat{k}} - \tilde{\Omega}^{\text{op}}_{k^*}\|^2_{\text{op}} \leq C_L\frac{\log p + k^*}{n} \leq C n^{-\frac{2\alpha-1}{2\alpha}} + C\frac{\log p}{n}. \tag{2.28}$$

It suffices to show that $\hat{k} \leq k^*$ with high probability. The following lemma, a probability version of Theorem 3, facilitates our proof of this claim.

Lemma 15. *Assume $\lceil n^{\frac{1}{2\alpha}} \rceil \leq p$. Then for any constant $C_1 > 0$, there exists a sufficiently large constant $C > 0$ irrelevant of α such that the local cropping estimator defined in (2.6) satisfies*

$$\mathbb{P}(\|\tilde{\Omega}^{\text{op}}_k - \Omega\|^2_{\text{op}} \leq C k^{-2\alpha+1} + C\frac{\log p + k}{n}) > 1 - \exp(-C_1(\log p + k)),$$

simultaneously for each $k \in \mathcal{H}$ and each $\Omega \in \mathcal{P}_\alpha(\eta, M)$ with $\alpha > \frac{1}{2}$.

Notice that for any l, we have $\|\tilde{\Omega}^{\text{op}}_{k^*} - \tilde{\Omega}^{\text{op}}_l\|^2_{\text{op}} \leq 2\|\tilde{\Omega}^{\text{op}}_{k^*} - \Omega\|^2_{\text{op}} + 2\|\Omega - \tilde{\Omega}^{\text{op}}_l\|^2_{\text{op}}$. Thus,

$$\begin{aligned}
&\mathbb{P}(\hat{k} > k^*) \\
\leq& \sum_{l \geq k^*} \mathbb{P}(\|\tilde{\Omega}^{\text{op}}_{k^*} - \tilde{\Omega}^{\text{op}}_l\|^2_{\text{op}} > C_L\frac{\log p + l}{n}) \\
\leq& \sum_{l \geq k^*} \Big(\mathbb{P}(\|\tilde{\Omega}^{\text{op}}_{k^*} - \Omega\|^2_{\text{op}} > \frac{C_L}{4}\frac{\log p + k^*}{n}) + \mathbb{P}(\|\tilde{\Omega}^{\text{op}}_l - \Omega\|^2_{\text{op}} > \frac{C_L}{4}\frac{\log p + l}{n})\Big) \\
\leq& n\Big(\exp\big(-C_1(\log p + k^*)\big) + \exp\big(-C_1(\log p + l)\big)\Big) \\
\leq& n^{-1}\eta^{-2}.
\end{aligned} \tag{2.29}$$

We have used the fact $k^* \leq l$ and the definition of k^* in the inequalities above, noting that a sufficiently large $C_1 > 0$ can be picked to guarantee the last inequality holds. The second to last inequality holds because of Lemma 15 and a sufficiently large C_L. Therefore, we have shown that the event $\hat{k} \leq k^*$ holds with probability at least $1 - n^{-1}\eta^{-2}$.

In the end, combining (2.26)-(2.29), we obtain that for any $\Omega \in \mathcal{P}_\alpha(\eta, M)$,

$$
\begin{aligned}
&\mathbb{E}\|\tilde{\Omega}_{\hat{k}}^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 \\
\leq &2\mathbb{E}\|\tilde{\Omega}_{k^*}^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 + 2\mathbb{E}\big(\|\tilde{\Omega}_{\hat{k}}^{\mathrm{op}} - \tilde{\Omega}_{k^*}^{\mathrm{op}}\|_{\mathrm{op}}^2 : \hat{k} \leq k^*\big) + 2\mathbb{E}\big(\|\tilde{\Omega}_{\hat{k}}^{\mathrm{op}} - \tilde{\Omega}_{k^*}^{\mathrm{op}}\|_{\mathrm{op}}^2 : \hat{k} > k^*\big) \\
\leq &Cn^{-\frac{2\alpha-1}{2\alpha}} + C\frac{\log p}{n} + 8\eta^2\mathbb{P}(\hat{k} > k^*) \\
\leq &Cn^{-\frac{2\alpha-1}{2\alpha}} + C\frac{\log p}{n} + 8n^{-1} \\
\leq &C(n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n}),
\end{aligned}
$$

where we also used that $\|\tilde{\Omega}_{\hat{k}}^{\mathrm{op}} - \tilde{\Omega}_{k^*}^{\mathrm{op}}\|_{\mathrm{op}}^2 \leq 4\eta^2$ in the second inequality. Therefore, we complete the proof. □

2.4 AN EXTENSION TO NONPARANORMAL DISTRIBUTIONS

In this section, we extend the minimax framework to the nonparanormal model. Assume that $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ follows the p-variate Gaussian distribution with covariance matrix Σ. Instead of n i.i.d. copies $\mathbf{X}_1, \mathbf{X}_1, \ldots, \mathbf{X}_n$ of $\mathbf{X}$, we only observe their transformations. Specifically, we denote the transformed variables of $\mathbf{X}$ by $\mathbf{Y} = (f_1(X_1), f_2(X_2), \ldots, f_p(X_p))^T$, where each f_i is some unknown strictly increasing function. Then our observation is $\mathbf{Z} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)^T \in \mathbb{R}^{n\times p}$, where each $\mathbf{Y}_i$ is the transformed $\mathbf{X}_i$. This is a form of the Gaussian copula model [Bickel et al., 1993], or the nonparanormal model [Liu et al., 2009]. To avoid the identifiability issue, we set $\mathrm{diag}(\Sigma) = I$, which makes Σ the correlation matrix. Here we consider the same structural assumption as in previous sections on the inverse of the correlation matrix, which is denoted by Ω. Based on $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ defined in

(1.1.5.1) and (1.1.5.1), the following two types of parameter spaces are of interest,

$$\mathcal{P}'_\alpha(\eta, M) = \left\{ \{\Omega, \{f_i\}\} : \begin{array}{c} \operatorname{diag}(\Omega^{-1}) = I, \quad \Omega \in \mathcal{P}_\alpha(\eta, M); \\ f_i \text{ is strictly increasing, } i \in [p]. \end{array} \right\}, \tag{2.30}$$

and

$$\mathcal{Q}'_\alpha(\eta, M) = \left\{ \{\Omega, \{f_i\}\} : \begin{array}{c} \operatorname{diag}(\Omega^{-1}) = I, \quad \Omega \in \mathcal{Q}_\alpha(\eta, M); \\ f_i \text{ is strictly increasing, } i \in [p]. \end{array} \right\}. \tag{2.31}$$

Our goal is to estimate the latent correlation structure, the inverse of the correlation matrix Ω, using the observation $\mathbf{Z}$. We establish the minimax risk of estimating Ω over the parameter spaces $\mathcal{P}'_\alpha(\eta, M)$ and $\mathcal{Q}'_\alpha(\eta, M)$ under the operator norm in the following theorem.

Theorem 8. *Assume $\log p = O(n)$, $n = O(p)$. Then for the nonparanormal model, the minimax risk of estimating Ω under the operator norm over $\mathcal{P}'_\alpha(\eta, M)$ with $\alpha > \frac{1}{2}$ satisfies*

$$\inf_{\tilde{\Omega}} \sup_{\{\Omega, \{f_i\}\} \in \mathcal{P}'_\alpha(\eta, M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2 \asymp n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n}. \tag{2.32}$$

The minimax risk of estimating Ω under the operator norm over $\mathcal{Q}'_\alpha(\eta, M)$ satisfies

$$\inf_{\tilde{\Omega}} \sup_{\{\Omega, \{f_i\}\} \in \mathcal{Q}'_\alpha(\eta, M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2 \asymp n^{-\frac{2\alpha}{1+2\alpha}} + \frac{\log p}{n}. \tag{2.33}$$

Rank-based estimator are widely applied in the nonparanormal model. Progress has been made in this field during the last decade especially for high-dimensional statistics. For instance, see Mitra and Zhang [2014] for bandable correlation matrix estimation, [Barber and Kolar, 2015] for Gaussian graphical models, and Fan et al. [2016] for multi-task regression via Cholesky decomposition.

2.4.1 Estimation Procedure

We introduce our rate-optimal estimation procedure over the parameter spaces $\mathcal{P}'_\alpha(\eta, M)$ and $\mathcal{Q}'_\alpha(\eta, M)$ under the operator norm. The approach to estimate the inverse of the correlation matrix in nonparanormal model is almost the same as the estimation scheme of the precision matrix under the operator norm in Section 2.1, except that the sample covariance matrix needs to be replaced by its rank-based nonparametric variant via Kendall's tau (τ) Kendall [1938] or Spearman's correlation coefficient rho (ρ) Spearman [1904].

Kendall's tau is defined as

$$\hat{\tau}_{ij} = \frac{2}{n(n-1)} \sum_{1 \le k_1 < k_2 \le n} \mathsf{sgn}(Z_{k_1 i} - Z_{k_2 i}) \mathsf{sgn}(Z_{k_1 j} - Z_{k_2 j}).$$

Then define

$$\hat{\Sigma}^\tau = [\sin(\frac{\pi}{2}\hat{\tau}_{ij})]_{p \times p}. \tag{2.34}$$

Spearman's rho is defined as

$$\hat{\rho}_{ij} = \frac{\sum_{k=1}^n (r_{ki} - (n+1)/2)(r_{kj} - (n+1)/2)}{\sqrt{\sum_{k=1}^n (r_{ki} - (n+1)/2)^2 \sum_{k=1}^n (r_{kj} - (n+1)/2)^2}},$$

where r_{ij} is the rank of Z_{ij} among $Z_{1j}, Z_{2j}, \ldots, Z_{nj}$. Define

$$\hat{\Sigma}^\rho = [2\sin(\frac{\pi}{6}\hat{\rho}_{ij})]_{p \times p}. \tag{2.35}$$

It is well-known that both $\hat{\Sigma}^\tau$ and $\hat{\Sigma}^\rho$ are unbiased estimators of the population correlation matrix Σ. We adopt almost the same estimation procedure proposed in Section 2.1, but replacing $\frac{1}{n}\mathbf{Z}^T\mathbf{Z}$ in (2.4) with either $\hat{\Sigma}^\tau$ or $\hat{\Sigma}^\rho$. In this way, we construct the nonparametric local cropping estimators $\tilde{\Omega}_k^\tau$ and $\tilde{\Omega}_k^\rho$ in replace of $\tilde{\Omega}_k^{\mathrm{op}}$ in (2.6). Note that the optimal choices of the bandwidth k are picked differently over two types of parameter spaces $\mathcal{P}'_\alpha(\eta, M)$ and $\mathcal{Q}'_\alpha(\eta, M)$ as we did over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ in Section 2.1.

2.4.2 Minimax Optimality

In this section, we prove that both $\tilde{\Omega}_k^\tau$ and $\tilde{\Omega}_k^\rho$ achieve the minimax optimality under the operator norm over the parameter space $\mathcal{P}'_\alpha(\eta, M)$. The minimax optimality over $\mathcal{Q}'_\alpha(\eta, M)$ can be established in the same way.

First, we derive the risk lower bound over $\mathcal{P}'_\alpha(\eta, M)$. Following the same strategy in Section 2.2.2, we construct the two subset $\mathcal{P}'_1$ and $\mathcal{P}'_2$ based on $\mathcal{P}_1$ and $\mathcal{P}_2$ given in (2.16) and (2.17). Define the subset

$$\mathcal{P}'_1 = \left\{ \{\Omega, \{f_i\}\} : \begin{array}{ll} \Omega = \mathrm{diag}(\Omega'^{-1})^{\frac{1}{2}} \Omega' \mathrm{diag}(\Omega'^{-1})^{\frac{1}{2}}, & \Omega' \in \mathcal{P}_1; \\ f_i(x) = \mathrm{diag}(\Omega'^{-1})_i^{\frac{1}{2}} x, & i \in [p]. \end{array} \right\}, \tag{2.36}$$

Lemma 16. *$\mathcal{P}'_1$ is a subset of $\mathcal{P}'_\alpha(\eta^2, M\eta)$.*

One can easily check that the probability measure of the Gaussian distribution with precision matrix Ω and transformation $\{f_i\}$, where $\{\Omega, \{f_i\}\} \in \mathcal{P}'_1$ is equivalent to the the probability measure of the Gaussian distribution with precision matrix Ω', where $\Omega' \in \mathcal{P}_1$. Therefore, by this one-to-one correspondence of probability measure between index sets $\mathcal{P}'_1$ and $\mathcal{P}_1$, we immediately have

$$\sup_{\{\Omega, \{f_i\}\} \in \mathcal{P}'_\alpha(\eta, M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2 \geq \sup_{\{\Omega, \{f_i\}\} \in \mathcal{P}'_1} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2 = \sup_{\Omega' \in \mathcal{P}_1} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2. \tag{2.37}$$

Applying the same proof strategy in Section 2.2.2, we have the following result.

Lemma 17. *We set $\Omega = \mathrm{diag}(\Omega'^{-1})^{\frac{1}{2}} \Omega' \mathrm{diag}(\Omega'^{-1})^{\frac{1}{2}}$ for each $\Omega'(\theta) \in \mathcal{P}_1$ in (2.16). Then it holds that*

$$\min_{H(\theta, \theta') \geq 1} \frac{\|\Omega(\theta) - \Omega(\theta')\|_2^2}{H(\theta, \theta')} \geq (\tau n^{-\frac{1}{2}})^2,$$

where $0 < \tau < \min\{M\eta, \frac{1}{4}\eta^{-2}, \eta - 1\}$.

Then by Assoaud's lemma, we obtain that

$$\inf_{\tilde{\Omega}} \sup_{\Omega' \in \mathcal{P}_1} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2 \geq \frac{\tau^2}{32} n^{-\frac{2\alpha - 1}{2\alpha}}.$$

We can use a similar strategy to construct $\mathcal{P}'_2$. Note that in this case we need to put $\tau a^{\frac{1}{2}}$ on the first sub-diagonal in $I - A$, instead of the diagonal of Σ. Combined with the result from the Le Cam's lemma on the subset $\mathcal{P}'_2$, we have

$$\sup_{\{\Omega,\{f_i\}\}\in\mathcal{P}'_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \geq C(n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n}). \tag{2.38}$$

Next, we turn to the risk upper bound of our rank-based local cropping estimators. The risk can be decomposed into the bias terms and the variance term in the same fashion in Section 2.2.1. Since the bias terms are deterministic and only due to the bandable structure of the Cholesky factor of the inverse correlation matrix, the upper bounds of two bias terms we derived in Lemma 2 and Lemma 4 still hold. For the variance term, a simple extension of Theorem 1 in [Mitra and Zhang, 2014] provides the following result.

Lemma 18. *For any Ω such that diag$(\Omega^{-1}) = I$ and $\Omega \in \mathcal{P}_\alpha(\eta, M) \cup \mathcal{Q}_\alpha(\eta, M)$, we have*

$$\mathbb{E}\big(\max_{m\in[p]} \|(\mathbf{C}^{3k}_{m-k}(\hat{\Sigma}^\tau) - \mathbf{C}^{3k}_{m-k}(\Omega^{-1})\|_{\text{op}}^2\big) \leq C\frac{\log p + k}{n},$$

$$\mathbb{E}\big(\max_{m\in[p]} \|(\mathbf{C}^{3k}_{m-k}(\hat{\Sigma}^\rho) - \mathbf{C}^{3k}_{m-k}(\Omega^{-1})\|_{\text{op}}^2\big) \leq C\frac{\log p + k}{n}.$$

Replacing Lemma 3 by Lemma 18 and following the rest of the proof in Theorem 3, we finally obtain that

$$\sup_{\{\Omega,\{f_i\}\}\in\mathcal{P}'_\alpha(\eta,M)} (\mathbb{E}\|\tilde{\Omega}^\tau_k - \Omega\|_{\text{op}}^2 + \mathbb{E}\|\tilde{\Omega}^\rho_k - \Omega\|_{\text{op}}^2) \leq Cn^{-\frac{2\alpha-1}{2\alpha}} + C\frac{\log p}{n}. \tag{2.39}$$

The lower bound (2.38) and upper bound (2.39) together give the optimal rate of convergence (2.32) in Theorem 8. The optimal rate of convergence (2.33) in Theorem 8 can be proved similarly. Therefore, we establish the minimax framework for nonparanormal distributions.

3.0 METHODOLOGY AND OPTIMALITY UNDER THE FROBENIUS NORM

3.1 ESTIMATION PROCEDURE

Under the Frobenius norm, our estimation procedure is based on the Cholesky decomposition of the precision matrix (1.1). More specifically, we estimate the matrix A and D respectively by auto-regression, and then combine them together to construct the estimator of Ω. The following estimation procedure applies to both the parameter space $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ as we will show that they enjoy the same optimal rate of convergence in Section 3.2.

Our estimator of the i-th row of A is based on the regression of X_i against its previous variables. Unlike those existing methods (Wu and Pourahmadi [2003], Huang et al. [2006], Bickel and Levina [2008b]) which rely on certain banding or penalized approaches for such a regression problem, we apply a block-thresholding procedure due to the decay structure in $\mathcal{P}_\alpha(\eta, M)$ which is defined in terms of nested ℓ_1 norm. To this end, we first regress X_i against $\mathbf{X}_{i-k_1:i-1} = (X_{i-k_1}, \ldots, X_{i-1})^T$ with bandwidth $k_1 = \lceil \frac{n}{c} \rceil$ with some sufficiently large $c > 0$. Recall that the $n \times 1$ matrix $\mathbf{Z}_i$ consists of n observations of X_i, and the $n \times k_1$ matrix $\mathbf{Z}_{i-k_1:i-1}$ represents n observations of $\mathbf{X}_{i-k_1:i-1}$. The empirical regression coefficients are

$$(\hat{a}_{i(i-k_1)}, \ldots \hat{a}_{i(i-1)})^T = (\mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_{i-k_1:i-1})^{-1} \mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_i. \tag{3.1}$$

We then further threshold the coefficients by taking advantages of the bandable structure of the Cholesky factor A. Specifically, we define $\hat{\mathbf{a}}_i^* \in \mathbb{R}^{i-1}$ with coordinate $\hat{a}_{ij}^*$ as follows,

$$\hat{a}_{ij}^* = \begin{cases} \hat{a}_{ij}, & \text{if } i - k_0 < j \le i-1, \\ \hat{a}_{ij}\mathbf{1}\left(|\hat{a}_{ij}| > \lambda_j\right), & \text{if } i-k_1 < j \le i-k_0, \\ 0, & \text{if } 1 \le j \le i-k_1, \end{cases} \tag{3.2}$$

where $k_0 = \lceil n^{\frac{1}{2\alpha+2}} \rceil$, $\lambda_j = (\lceil \log_2^{i-j} - \log_2^{k_0} \rceil R)^{\frac{1}{2}}$, and $R = 8\eta \|(\mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\text{op}}$. Note that we keep the last k_0 coefficients and apply a block-thresholding rule in which the size of each block doubles backwards sequentially for the remaining coefficients in (3.2). Our procedure is inspired by the optimal estimation procedure over Besov balls for many nonparametric function estimation problems, or equivalently, the corresponding Gaussian sequence models (See Cai [2012] the reference therein). We emphasize that any linear estimator of the coefficients $(\hat{a}_{i(i-k_1)}, \ldots \hat{a}_{i(i-1)})^T$ cannot yield to the optimal rates of convergence in our setting under the Frobenius norm.

Our estimator of $I - A$ can be constructed by arranging zero-padded $\hat{\mathbf{a}}_i^{*T}$, $i \in [p]$ accordingly with an identity matrix. Specifically, set the ij-th entry of $\hat{A}^*$ as $\hat{a}_{ij}^*$ when $i \in [p]$, $j \in [i-1]$, otherwise as zero. We also need to bound the singular values of $(I - \hat{A}^*)$. To this end, we define

$$\widetilde{I-A} = \mathbf{P}_\eta(I - \hat{A}^*),$$

as our estimator of $(I - A)$, where $\mathbf{P}_\eta(\cdot)$ is defined in (2.3).

The estimation of D is based on the sample variances of those empirical residuals in the regression of X_i against $\mathbf{X}_{i-k_1:i-1} = (X_{i-k_1}, \ldots, X_{i-1})^T$. For each i, the sample variance of the empirical residual is

$$\hat{d}_i = \frac{1}{n-k_1} \mathbf{Z}_i^T (I - \mathbf{M}_i)^T (I - \mathbf{M}_i) \mathbf{Z}_i, \tag{3.3}$$

where $\mathbf{M}_i = \mathbf{Z}_{i-k_1:i-1}(\mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_{i-k_1:i-1})^{-1} \mathbf{Z}_{i-k_1:i-1}^T$.

Let $\hat{D} = \text{diag}(\hat{d})$, where $\hat{d} = (\hat{d}_1, \ldots, \hat{d}_p)^T$. We define $\tilde{D} = \mathbf{P}_\eta(\hat{D})$ as our estimator of D. Finally, define our estimator of Ω as

$$\tilde{\Omega}_k^{\text{F}} = (\widetilde{I-A})^T \tilde{D}^{-1} (\widetilde{I-A}). \tag{3.4}$$

Remark 6. *For the parameter space $\mathcal{Q}_\alpha(\eta, M)$, a much simpler banding estimation scheme on the empirical regression coefficients is able to achieve the minimax risk. Set $k = \lceil n^{\frac{1}{2\alpha+2}} \rceil$. We use the empirical residuals and coefficients obtained by regressing each X_i against $\mathbf{X}_{i-k:i-1}$ to directly construct the estimators of A and D. It can be proved that this estimator achieves the minimax risk over the parameter space $\mathcal{Q}_\alpha(\eta, M)$.*

3.2 RATE OPTIMALITY UNDER THE FROBENIUS NORM

In this section, we establish that the optimal rates of convergence for estimating the precision matrix over the parameter spaces $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ are identical under the Frobenius norm. Since $\mathcal{Q}_\alpha(\eta, \alpha M) \subset \mathcal{P}_\alpha(\eta, M)$, one immediately obtain that

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2 \geq \inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,\alpha M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2. \tag{3.5}$$

In order to show (1.9) in Theorem 2, it suffices to establish the upper bound over the parameter space $\mathcal{P}_\alpha(\eta, M)$ and the matching lower bound over the parameter space $\mathcal{Q}_\alpha(\eta, M)$. We assume that $\mathbf{X}$ follows the p-variate Gaussian distribution, with mean zero and precision matrix Ω in this section.

3.2.1 Minimax Upper Bound under the Frobenius Norm

In this section, we establish the following risk upper bound of the regression-based block-thresholding estimation procedure we proposed in Section 3.1 under the Frobenius norm over $\mathcal{P}_\alpha(\eta, M)$.

Theorem 9. *Assume $\lceil n^{\frac{1}{2\alpha+2}} \rceil \leq p$. The estimator defined in (3.4) of the precision matrix Ω over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, \alpha M)$ given in (1.1.5.1) and (1.1.5.1) with $k = \lceil n^{\frac{1}{2\alpha+2}} \rceil$ satisfies*

$$\sup_{\mathcal{Q}_\alpha(\eta,\alpha M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega}_k^{\mathrm{F}} - \Omega\|_{\mathrm{F}}^2 \leq \sup_{\mathcal{P}_\alpha(\eta,M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega}_k^{\mathrm{F}} - \Omega\|_{\mathrm{F}}^2 \leq Cn^{-\frac{2\alpha+1}{2\alpha+2}}. \tag{3.6}$$

Proof. We focus on the second inequality since the first one is trivial. Note that $\tilde{\Omega}_k^{\mathrm{F}} = (\widetilde{I-A})^T \tilde{D}^{-1}(\widetilde{I-A})$ according to (3.4) while $\Omega = (I-A)^T D^{-1}(I-A)$. The risk upper bound can be controlled by bounding $\widetilde{I-A} - (I-A)$ and $\tilde{D} - D$. To this end, we first provide some properties of our estimator.

Lemma 19. *Assume that $\mathbf{X}$ follows the p-variate Gaussian distribution with mean zero and precision matrix $\Omega = (I-A)^T D^{-1}(I-A)$, which belongs to parameter space $\mathcal{P}_\alpha(\eta, M)$ defined in (1.1.5.1). For any fixed i, d_i is the i-th diagonal of D, $\mathbf{a}_i \in \mathbb{R}^{i-1}$ corresponds the i-th row of the lower triangle in A. $\hat{d}_i$ is defined in (3.3), and $\hat{\mathbf{a}}_i^* \in \mathbb{R}^{i-1}$ corresponds the i-th row of the lower triangle in $\hat{A}^*$ defined in (3.2). Then we have*

$$\mathbb{E}|\hat{d}_i - d_i|^2 \le Cn^{-\frac{2\alpha+1}{2\alpha+2}},$$
$$\mathbb{E}\|\hat{\mathbf{a}}_i^* - \mathbf{a}_i\|_2^2 \le Cn^{-\frac{2\alpha+1}{2\alpha+2}}.$$

We are ready to establish the upper bounds of $\widetilde{I-A} - (I-A)$ and $\tilde{D} - D$ separately. Note that $\|\tilde{D}^{-1}\|_{\mathrm{op}} \le \eta$ and $\|D^{-1}\|_{\mathrm{op}} \le \eta$, which is due to Lemma 24. Therefore, Lemma 23 yields $\mathbb{E}\|\tilde{D} - D\|_{\mathrm{F}}^2 \le 4\mathbb{E}\|\hat{D} - D\|_{\mathrm{F}}^2$, which further implies that

$$\begin{aligned}
\frac{1}{p}\mathbb{E}\|D^{-1} - \tilde{D}^{-1}\|_{\mathrm{F}}^2 &\le \frac{1}{p}\mathbb{E}\|\tilde{D}^{-1}\|_{\mathrm{op}}^2 \|\tilde{D} - D\|_{\mathrm{F}}^2 \|D^{-1}\|_{\mathrm{op}}^2 \\
&\le 4\eta^4 \frac{1}{p}\mathbb{E}\|\hat{D} - D\|_{\mathrm{F}}^2 \\
&\le 4\eta^4 \frac{1}{p}\sum_i \mathbb{E}|\hat{d}_i - d_i|^2.
\end{aligned}$$

Together with Lemma 19, it follows that

$$\frac{1}{p}\mathbb{E}\|D^{-1} - \tilde{D}^{-1}\|_{\mathrm{F}}^2 \le Cn^{-\frac{2\alpha+1}{2\alpha+2}}. \tag{3.7}$$

Next, we turn to prove that $\frac{1}{p}\mathbb{E}\|(\widetilde{I-A}) - (I-A)\|_{\mathrm{F}}^2 \le Cn^{-\frac{2\alpha+1}{2\alpha+2}}$. Lemma 23 implies

$$\frac{1}{p}\mathbb{E}\|(\widetilde{I-A}) - (I-A)\|_{\mathrm{F}}^2 \le \frac{4}{p}\mathbb{E}\|\hat{A}^* - A\|_{\mathrm{F}}^2 \le \frac{4}{p}\sum_i \mathbb{E}\|\hat{\mathbf{a}}_i^* - \mathbf{a}_i\|_2^2.$$

Combining above equation with Lemma 19, we have

$$\frac{1}{p}\mathbb{E}\|(\widetilde{I-A}) - (I-A)\|_{\mathrm{F}}^2 \le Cn^{-\frac{2\alpha+1}{2\alpha+2}}. \tag{3.8}$$

At last, we derive the risk upper bound of our estimator. It is clear that $\|\widetilde{I-A}\|_{\mathrm{op}} \leq \eta$, $\|\tilde{D}^{-1}\|_{\mathrm{op}} \leq \eta$. According to Lemma 24, $\|I-A\|_{\mathrm{op}} \leq \eta$, $\|D^{-1}\|_{\mathrm{op}} \leq \eta$. Combining these facts with (3.7) and (3.8), we have

$$\begin{aligned}
\frac{1}{p}\mathbb{E}\|\tilde{\Omega}-\Omega\|_{\mathrm{F}}^2 \leq & \frac{3}{p}\mathbb{E}\big(\|I-A\|_{\mathrm{op}}^2\|D^{-1}\|_{\mathrm{op}}^2\|(\widetilde{I-A})-(I-A)\|_{\mathrm{F}}^2 \\
& + \|I-A\|_{\mathrm{op}}^2\|D^{-1}-\tilde{D}^{-1}\|_{\mathrm{F}}^2\|\widetilde{I-A}\|_{\mathrm{op}}^2 \\
& + \|(\widetilde{I-A})-(I-A)\|_{\mathrm{F}}^2\|\tilde{D}^{-1}\|_{\mathrm{op}}^2\|\widetilde{I-A}\|_{\mathrm{op}}^2\big) \\
\leq & 6\eta^4\frac{1}{p}\mathbb{E}\|(\widetilde{I-A})-(I-A)\|_{\mathrm{F}}^2 + 3\eta^4\frac{1}{p}\mathbb{E}\|D^{-1}-\tilde{D}^{-1}\|_{\mathrm{F}}^2 \\
\leq & Cn^{-\frac{2\alpha+1}{2\alpha+2}}.
\end{aligned}$$

Therefore, we finish the proof of Theorem 9. □

3.2.2 Minimax Lower Bound under the Frobenius Norm

In this section, we establish the matching lower bound $n^{-\frac{2\alpha+1}{2\alpha+2}}$ over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$.

Theorem 10. *The minimax risk for estimating the precision matrix Ω over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, \alpha M)$ under the Frobenius norm satisfies*

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega}-\Omega\|_F^2 \geq \inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,\alpha M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega}-\Omega\|_F^2 \geq \frac{\tau^2}{32}n^{-\frac{2\alpha+1}{2\alpha+2}}.$$

Remark 7. *The minimax risk for estimating the precision matrices over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ under the Frobenius norm in Theorem 2 immediately follows from Theorems 9 and 10.*

Proof. It is sufficient to establish the lower bound over $\mathcal{Q}_\alpha(\eta, M)$ since the first inequality immediately follows from (3.5). We construct a least favorable subset in $\mathcal{Q}_\alpha(\eta, M)$. Without loss of generality, we assume $\frac{p}{2k}$ is an integer where $k = \min\{\lceil n^{\frac{1}{2\alpha+2}}\rceil, \frac{p}{2}\}$. Define the index set $\Theta' = \{0,1\}^{\frac{kp}{2}}$. For each $\theta \in \Theta'$, we further denote it as $\frac{p}{2k}$ many k^2 dimensional vectors, i.e., $\theta = \{\theta(s)\}_{1\leq s\leq \lceil\frac{p}{2k}\rceil}$, where $\theta(s)_{ij}$ is equal to 0 or 1. For such an index θ, there is a corresponding $p \times p$ block diagonal matrix $C(\theta)$ such that each $k \times k$ block $C_s(\theta(s)) \equiv$

$[c(s)_{ij}]_{k\times k}$, where $c(s)_{ij} = \tau n^{-\frac{1}{2}}\theta(s)_{ij}$, $s \in \lceil \frac{p}{2k} \rceil$. We set τ as a positive constant which is less than $\min\{M, \frac{1}{4}\eta^{-1}, \eta^{\frac{1}{2}} - 1\}$.

$$C(\theta) = \begin{bmatrix} \boxed{\begin{matrix} 0_k & 0_k \\ C_1(\theta(1)) & 0_k \end{matrix}} & 0_{2k} & \dots & 0_{2k} \\ 0_{2k} & \boxed{\begin{matrix} 0_k & 0_k \\ C_2(\theta(2)) & 0_k \end{matrix}} & \dots & 0_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{2k} & 0_{2k} & \dots & \boxed{\begin{matrix} 0_k & 0_k \\ C_{\lceil \frac{p}{2k} \rceil}(\theta(\lceil \frac{p}{2k} \rceil)) & 0_k \end{matrix}} \end{bmatrix}.$$

Finally, we define the subset of $\mathcal{Q}_\alpha(\eta, M)$ indexed by Θ' as follows

$$\mathcal{P}_4 = \left\{ \Omega(\theta) : \Omega(\theta) = (I_p - C(\theta))^T (I_p - C(\theta)), \theta \in \Theta' \right\}. \tag{3.9}$$

Lemma 20. *$\mathcal{P}_4$ is a subset of $\mathcal{Q}_\alpha(\eta, M)$.*

Applying Lemma 6 to $\mathcal{P}_4$, we obtain that,

$$\inf_{\tilde{\Omega}} \max_{\theta \in \Omega(\Theta')} 2^2 \mathbb{E}_\theta \|\tilde{\Omega} - \Omega(\theta)\|_F^2 \geq \min_{H(\theta,\theta') \geq 1} \frac{\|\Omega(\theta) - \Omega(\theta')\|_F^2}{H(\theta, \theta')} \frac{kp}{4} \min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\| \tag{3.10}$$

Lemma 21. *Let P_θ be the joint distribution of n i.i.d. observations from $N(0, \Omega(\theta)^{-1})$, where $\Omega(\theta) \in \mathcal{P}_4$ defined in (3.9). Then*

$$\min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\| \geq 0.5, \tag{3.11}$$

and

$$\min_{H(\theta,\theta') \geq 1} \frac{\|\Omega(\theta) - \Omega(\theta')\|_F^2}{H(\theta, \theta')} \geq \tau^2 n^{-1}. \tag{3.12}$$

Applying Lemma 21 into (3.10), we obtain

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \frac{1}{p} \mathbb{E}\|\tilde{\Omega} - \Omega\|_F^2 \geq \inf_{\tilde{\Omega}} \sup_{\mathcal{P}_4} \frac{1}{p} \mathbb{E}\|\tilde{\Omega} - \Omega\|_F^2 \geq \frac{\tau^2}{32} n^{-1} \min\left\{ n^{\frac{1}{2\alpha+2}}, \frac{p}{2} \right\},$$

which completes the proof of Theorem 10, noting that $n < p$. $\square$

4.0 NUMERICAL STUDIES

In this section, we turn to the numerical performance of the proposed rate-optimal estimators under the operator norm for $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ defined in (1.1.5.1) and (1.1.5.1) to further illustrate the fundamental difference of $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$. In addition, we compare them with the banding estimator proposed in [Bickel and Levina, 2008a], which is based on the auto-regression between variables. Specifically, for a given bandwidth $k < n$, the banding estimator is defined as $\tilde{\Omega}^{BL} = (I - \tilde{A}^{BL})^T(\tilde{D}^{BL})^{-1}(I - \tilde{A}^{BL})$. Here the i-th row of the lower triangular matrix $\tilde{A}^{BL}$ is the vector $\hat{\mathbf{a}}_i$ in (3.1), i.e., the least square estimates of the coefficients for the regression of X_i against $\mathbf{X}_{i-k:i-1}$. The i-th entry of the diagonal matrix $\tilde{D}^{BL}$ is the estimate of the residual variance for the regression of X_i against $\mathbf{X}_{i-k:i-1}$.

4.1 SIMULATION IN $\mathcal{Q}_\alpha(\eta, M)$ UNDER THE OPERATOR NORM

We first focus on the parameter space $\mathcal{Q}_\alpha(\eta, M)$ and compare the performance of local cropping estimator and the banding estimator. Specifically, we generate the precision matrix in the following form:

$$\Omega = (I - A)^T D^{-1}(I - A), \quad A \equiv [a_{ij}]_{p\times p}, \quad D = I_p,$$

where $a_{ij} = -(i-j)^{-\alpha-1}$ when $i > j$; otherwise $a_{ij} = 0$. It is easy to check that $\Omega \in \mathcal{Q}_\alpha(\eta, 1)$ with some large $\eta > 0$. The simulation is done with a range of parameter values for p, n, α. Specifically, the decay rate α ranges from 0.5 to 2 with a step of 0.5, the sample size n ranges from 500 to 4000, the dimension p ranges from 500 to 2000.

In this setting, we compare our local cropping estimator (denoted as cropping.Q.) with the banding estimator (denoted as BL) proposed in [Bickel and Levina, 2008a]. According to [Bickel and Levina, 2008a], the bandwidth of banding estimator is chosen as $k \asymp (n/\log p)^{1/(2\alpha+2)}$. The optimal bandwidth over $\mathcal{Q}_\alpha(\eta, M)$ is $k \asymp n^{1/(2\alpha+1)}$. In the simulation, the bandwidth of BL estimator is $\lfloor (n/\log p)^{1/(2\alpha+2)} \rfloor$ and the bandwidth of crop.Q is $\lfloor n^{1/(2\alpha+1)} \rfloor$.

Table 1 reports the average errors of the banding estimator (BL) and local cropping estimator (crop.Q) under the operator norm over 100 replications. The smaller errors in each experiment are highlighted in boldface. Figure 2 displays the boxplots of the errors of BL and crop.Q.

It can be seen from Table 1 that crop.Q outperforms BL in most cases with a few exceptions when n is small. As the sample size increases, the average errors of both methods decrease, which matches our intuition. In addition, the dimension p has minor effect on the errors of both estimators, which is partially reflected by the optimal rates (dominating term $n^{-\frac{2\alpha}{2\alpha+1}}$) obtained in Theorem 1. For each fixed dimension p, the superiority crop.Q over BL becomes more significant as the sample size n increases, which implies that BL estimator is indeed sub-optimal.

4.2 SIMULATION IN $\mathcal{P}_\alpha(\eta, M)$ UNDER THE OPERATOR NORM

We demonstrate the fundamental difference between two types of parameter space $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ by numerical studies in this section. Of note, although local cropping estimators proposed in (2.6) are rate-optimal over both $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$, the corresponding optimal choices of bandwidth are distinct. We generate precision matrices in the following way to guarantee that Ω is always in $\mathcal{P}_\alpha(\eta, M)$ but not in $\mathcal{Q}_\alpha(\eta, M)$ with some fixed η and M. Considering

$$\Omega = (I - A)^T D^{-1} (I - A), \quad A \equiv [a_{ij}]_{p\times p}, \quad D = I_p,$$

where the first column of A is $a_{i1} = -2(i-1)^{-\alpha}$, $2 \le i \le p$. The remaining entries are all zeros. It is easy to check that $\Omega \in \mathcal{P}_\alpha(\eta, 2)$ with some large $\eta > 0$. The simulation is carried out with a similar range of values for p, n, α as in Section 4.1. Note that the consistent estimator exists only if $\alpha > 0.5$. Therefore, in this setting, the decay rate α varies among 1, 1.5 and 2.

The optimal choice of bandwidth of local cropping estimator over $\mathcal{P}_\alpha(\eta, M)$ is $k \asymp n^{\frac{1}{2\alpha}}$, which is different from the one of crop.Q. We denote this rate-optimal estimator in $\mathcal{P}_\alpha(\eta, M)$ by crop.P. In the simulation, the bandwidth of crop.P is $\lfloor n^{\frac{1}{2\alpha}} \rfloor$. We also include BL estimator as a reference.

Table 2 reports the average errors of the three procedures, crop.P, crop.Q and BL, under the operator norm over 100 replications. The smallest errors in each experiment are highlighted in boldface. Figure 3 plots the boxplots of their errors for $p = 500, 1000, 2000$.

Since Ω always belongs to $\mathcal{P}_\alpha(\eta, M)$ but not $\mathcal{Q}_\alpha(\eta, M)$, the estimator crop.Q is sub-optimal and thus expected to have an inferior performance. Table 2 shows this point, i.e., for fixed p and α, the advantage of crop.P is more obvious as n increases. Especially, crop.P outperforms the other two estimators when $n = 4000$. We also see a similar pattern as in Table 1 that p has minor effect on the errors of all the estimators.

Theorem 1 immediately follows from Theorems 3, 4, 5 and 6 while Theorem 2 immediately follows from Theorems 9 and 10. We provide key lemmas in this section in the proofs of those main theorems. Recall that for $1 \le q \le \infty$, the matrix ℓ_q norm of a matrix S is defined by $\|S\|_q = \max_{\|x\|_q = 1} \|Sx\|_q$. We use the following two facts repeatedly in this section: (i) $\frac{\|S\|_{\mathrm{F}}}{\sqrt{p}} \le \|S\|_{\mathrm{op}} \le \sqrt{\|S\|_1 \|S\|_\infty}$; (ii) for a real symmetric matrix S, $\|S\|_{\mathrm{op}} \le \|S\|_1$.

Table 1: The average errors under the operator norm of the banding estimator (BL) and the local cropping estimator (crop.Q) over 100 replications.

p	n	$\alpha = 0.5$		$\alpha = 1$		$\alpha = 1.5$		$\alpha = 2$	
		crop.Q	BL	crop.Q	BL	crop.Q	BL	crop.Q	BL
500	500	**4.68**	5.44	**1.64**	2.38	1.18	**1.16**	0.93	**0.81**
	1000	**3.29**	4.89	**1.17**	1.72	**0.82**	1.08	**0.66**	0.69
	2000	**2.47**	4.45	**0.89**	1.33	**0.59**	0.69	**0.48**	0.59
	4000	**1.84**	3.80	**0.62**	1.07	**0.41**	0.64	**0.34**	0.53
1000	500	**4.96**	5.74	**1.75**	2.40	1.30	**1.19**	0.99	**0.84**
	1000	**3.43**	5.19	**1.24**	1.74	**0.86**	1.10	**0.68**	0.70
	2000	**2.58**	4.75	**0.93**	1.35	**0.62**	0.71	**0.51**	0.60
	4000	**1.93**	4.10	**0.66**	1.33	**0.44**	0.65	**0.36**	0.55
2000	500	**5.14**	5.97	**1.85**	2.41	1.33	**1.21**	1.06	**0.89**
	1000	**3.58**	5.41	**1.30**	1.76	**0.90**	1.12	0.72	**0.71**
	2000	**2.69**	4.97	**0.98**	1.37	**0.65**	0.73	**0.54**	0.62
	4000	**2.01**	4.32	**0.69**	1.34	**0.45**	0.66	**0.38**	0.55

Table 2: The average errors under the operator norm of the banding estimator (BL) and the local cropping estimator (crop.P & crop.Q) with two different bandwidths over 100 replications.

p	n	$\alpha = 1$			$\alpha = 1.5$			$\alpha = 2$		
		crop.P	crop.Q	BL	crop.P	crop.Q	BL	crop.P	crop.Q	BL
500	500	1.50	**1.18**	2.32	**0.66**	0.73	0.86	**0.52**	0.65	0.53
	1000	1.09	**0.96**	1.80	**0.47**	0.56	0.83	**0.38**	0.56	0.45
	2000	0.83	**0.80**	1.53	**0.35**	0.43	0.55	**0.27**	0.32	0.41
	4000	**0.64**	0.68	1.33	**0.26**	0.35	0.54	**0.19**	0.24	0.38
1000	500	1.50	**1.20**	2.36	**0.68**	0.74	0.91	**0.57**	0.68	0.59
	1000	1.12	**0.98**	1.82	**0.49**	0.58	0.81	**0.39**	0.55	0.46
	2000	0.84	**0.81**	1.54	**0.37**	0.44	0.55	**0.27**	0.32	0.41
	4000	**0.65**	0.68	1.52	**0.26**	0.35	0.53	**0.19**	0.24	0.38
2000	500	1.51	**1.21**	2.39	**0.69**	0.75	0.96	**0.62**	0.71	0.63
	1000	1.16	**1.00**	1.81	**0.51**	0.60	0.84	**0.39**	0.56	0.46
	2000	0.85	**0.81**	1.55	**0.39**	0.44	0.56	**0.27**	0.33	0.41
	4000	**0.65**	0.69	1.70	**0.26**	0.35	0.53	**0.19**	0.24	0.38

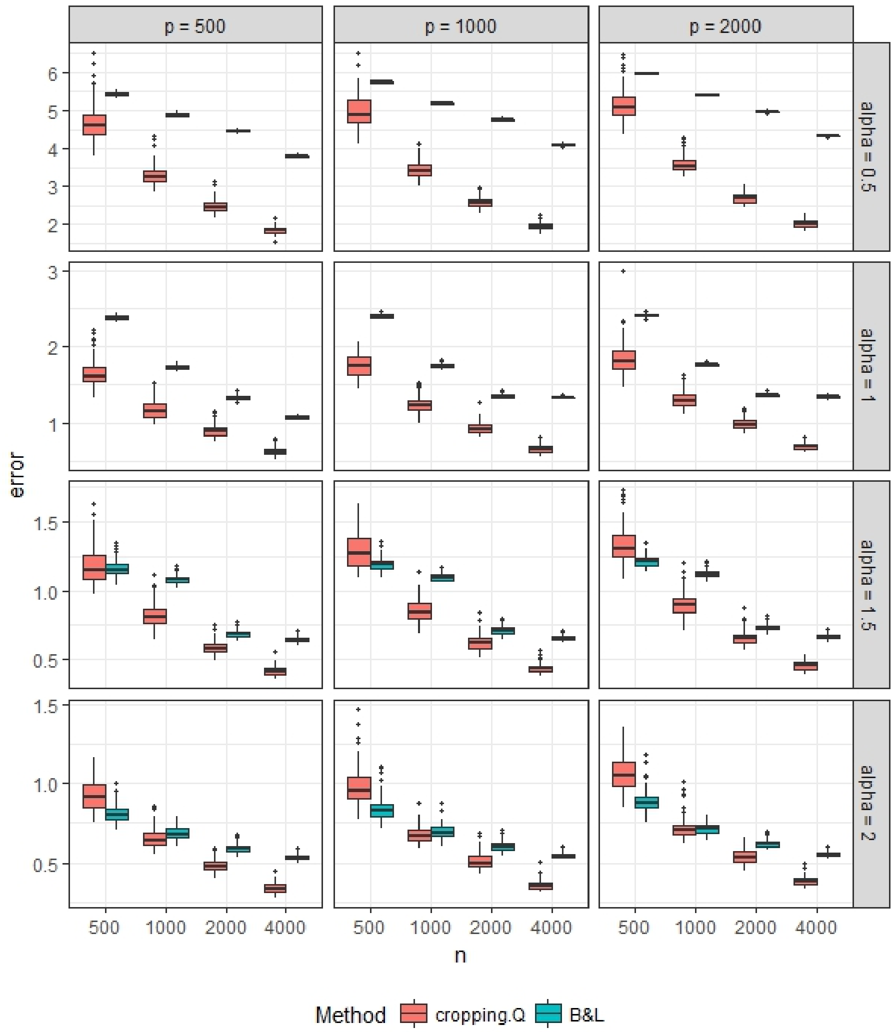

Figure 2: The boxplot of the errors from the local cropping estimator with the optimal bandwidth in $\mathcal{Q}_\alpha(\eta, M)$ (cropping.Q) and the banding estimator (BL) over 100 replications.

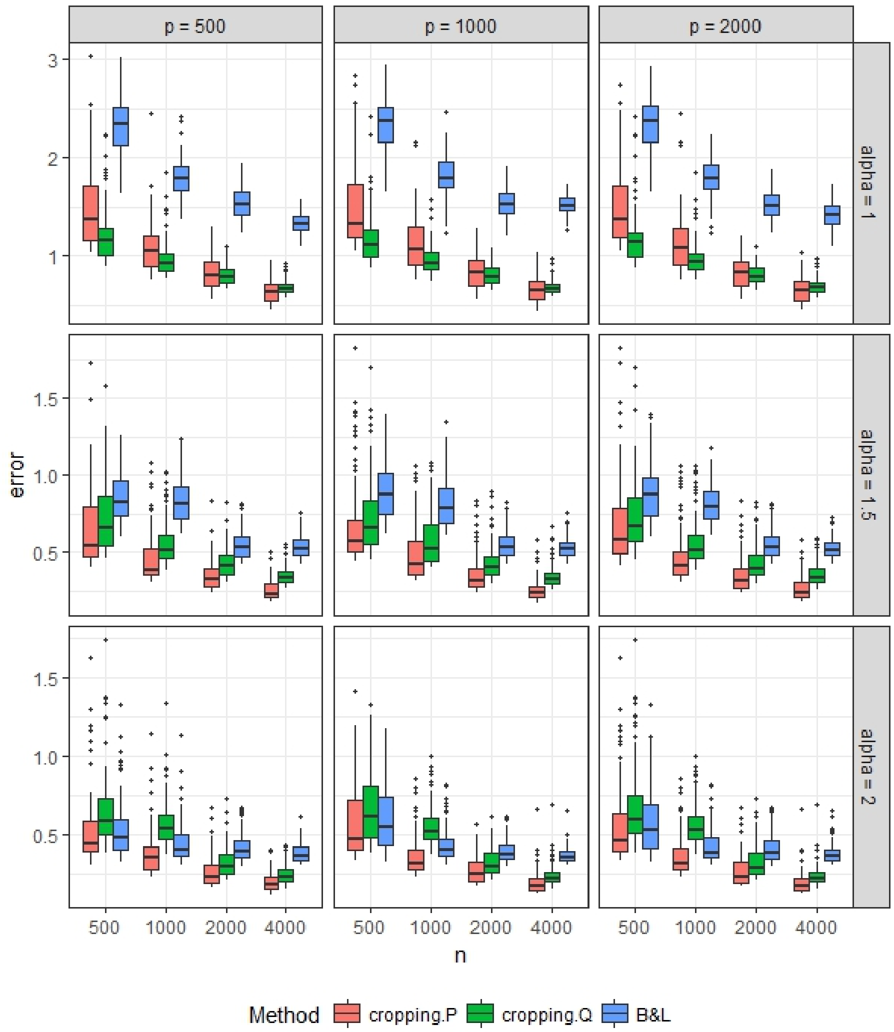

Figure 3: The boxplot of the errors from the local cropping estimator with the optimal bandwidth in $\mathcal{P}_\alpha(\eta, M)$ (cropping.P), the local cropping estimator with the optimal bandwidth in $\mathcal{Q}_\alpha(\eta, M)$ (cropping.Q) and the banding estimator (BL) over 100 replications.

APPENDIX A

DISCUSSION

Various structures on precision matrices besides the structure we discussed in this paper, precision matrices with bandable Cholesky factor, have been proposed in recent years. During the finalizing process of this paper, a similar structure was discussed by Hu and Negahban [2017], where the bandable structure is imposed on the entire precision matrix in contrast to the Cholesky factor. More specifically, a parameter space of bandable precision matrices was defined as follows,

$$\mathcal{F}_\alpha(\eta, M) = \Big\{\Omega : \quad 0 < \eta^{-1} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) < \eta, \\ \max_j \sum_{|i-j|\geq k} \omega_{ij} \leq Mk^{-\alpha}, \quad k \in [p]\Big\}.$$

In their paper, Hu and Negahban also established the minimax rate of convergence under the operator norm loss for this parameter space as follows,

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{F}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}.$$

Although the structure of the parameter space $\mathcal{F}_\alpha(\eta, M)$ above looks similar to that of $\mathcal{P}_\alpha(\eta, M)$ considered in our paper, there are fundamental differences in terms of interpretation as well as rates of convergence, which deserve a brief discussion. First of all, note that the interpretation of the entry in the precision matrix and the one in its Cholesky factor are different: ω_{ij} in Ω represents the partial covariance of X_i and X_j conditioned on the rest of variables, while a_{ij} in A represents the partial covariance of X_i and X_j conditioned on the variables whose index is smaller than j (assuming $i < j$). As a result, it is more common to consider parameter spaces $\mathcal{P}_\alpha(\eta, M)$ when all variable have a natural order as is the case in auto-regressive processes while the parameter space $\mathcal{F}_\alpha(\eta, M)$ is often used to capture a sense of locality among all variables. In addition, the minimax rate over $\mathcal{P}_\alpha(\eta, M)$ under the operator norm loss is $n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n}$, which is distinguished with the one $n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}$ over $\mathcal{F}_\alpha(\eta, M)$. We provide an example to partially explain this difference. Let $A \equiv [a_{ij}]_{p\times p}$ with $a_{i1} = i^{-\alpha}$ for $i > 1$, otherwise $a_{ij} = 0$. One can easily verify that for all dimension p, $\Omega = (I - A)^T(I - A)$ belongs to $\mathcal{P}_\alpha(\eta, 1)$ with some constant $\eta > 0$ but is not always in the space $\mathcal{F}_\alpha(\eta, M)$ with any fixed $M > 0$.

A more careful investigation on those two parameter spaces $\mathcal{F}_\alpha(\eta, M)$, $\mathcal{P}_\alpha(\eta, M)$, as well as the one $\mathcal{Q}_\alpha(\eta, M)$ considered in our paper reveals an interesting connection. On the one hand, we have that $\mathcal{Q}_\alpha(\eta, M) \subset \mathcal{F}_\alpha(\eta, CM)$. To see this, we need some facts in the proof of Lemma 11. According to (E.1) and (E.2) in Lemma 11, for any $\Omega \in \mathcal{Q}_\alpha(\eta, M)$, we have $\|bd_k(\Omega) - \Omega\|_1 \leq \|U_h\|_1 + \|U_h\|_\infty \leq CMk^{-\alpha}$ for any k, which immediately follows $\mathcal{Q}_\alpha(\eta, M) \subset \mathcal{F}_\alpha(\eta, CM)$, where $bd_h(\cdot)$ is defined in Equation (C.1). On the other hand, we can show that $\mathcal{F}_\alpha(\eta, M) \subset \mathcal{P}_\alpha(\eta, M')$ when $\alpha > 1$ and M is sufficiently small, which is summarized in the following lemma.

Lemma 22. *Assume $\alpha > 1$ and M is sufficiently small (depending on α), then we have $\mathcal{F}_\alpha(\eta, M) \subset \mathcal{P}_\alpha(\eta, M')$, where constant M' depends on M and η only.*

Proof. To facilitate the proof, we restate the definition of $\mathcal{F}_\alpha(\eta, M)$ and $\mathcal{P}_\alpha(\eta, M)$. Define a general bandable matrix space $\mathcal{L}_\alpha(M)$,

$$\mathcal{L}_\alpha(M) = \Big\{X : \max_j \sum_{|i-j|\geq k} |x_{ij}| \leq Mk^{-\alpha}, \quad k \in [p]\Big\}.$$

It is clear that $X \in \mathcal{F}_\alpha(\eta, M)$ is equivalent to that $0 < \eta^{-1} \leq \lambda_{\min}(X) \leq \lambda_{\max}(X) < \eta$ and $X \in \mathcal{L}_\alpha(M)$, while $X \in \mathcal{P}_\alpha(\eta, M)$ is equivalent to that $0 < \eta^{-1} \leq \lambda_{\min}(X) \leq \lambda_{\max}(X) < \eta$ and $A \in \mathcal{L}_\alpha(M)$, where $X = (I - A)^T D^{-1}(I - A)$. Recall that Lemma 24 shows that $\eta^{-1} \leq \lambda_{\min}(X) \leq \lambda_{\max}(X) < \eta$ implies that $\eta^{-1} \leq \lambda_{\min}(D) \leq \lambda_{\max}(D) \leq \eta$. Therefore, we have that $(I - A)^T(I - A) \in \mathcal{L}_\alpha(\eta M)$ whenever $X \in \mathcal{F}_\alpha(\eta, M)$. Then it suffices to prove that $(I - A)^T(I - A) \in \mathcal{L}_\alpha(M') \Rightarrow A \in \mathcal{L}_\alpha(M')$.

We use induction to show above claim. Define $\mathrm{R}_i(A)$ as the p by p matrix, which keeps the i-th row of the lower triangular matrix A and sets the rest zero. Similarly, we define $\mathrm{R}_{i:j}(A)$ as the p by p matrix, which keeps all rows between the i-th row and the j th row (inclusive) of A. To prove $A \in \mathcal{L}_\alpha(M')$, it is enough to show for any i, $\mathrm{R}_i(A) \in \mathcal{L}_\alpha(M')$.

We first show that $\mathrm{R}_p(A) \in \mathcal{L}_\alpha(M')$. Note that A is a lower triangular matrix with zeros on the diagonal. Therefore, all entries in the last row of $A^T A$ are equal to zero, which means,

$$\mathrm{R}_p((I-A)^T(I-A)) = \mathrm{R}_p(I) - \mathrm{R}_p(A).$$

Since $(I-A)^T(I-A) \in \mathcal{L}_\alpha(M')$, we obtain that $\mathrm{R}_p(A) \in \mathcal{L}_\alpha(M')$.

Second, suppose $\mathrm{R}_i(A) \in \mathcal{L}_\alpha(M')$ for all $i \in k+1 : p$, we show that $\mathrm{R}_k(A) \in \mathcal{L}_\alpha(M')$. We denote the first $k-1$ columns of a matrix B by $\mathrm{col}_{1:k-1}(B)$. By simple algebra, we have

$$\mathrm{col}_{1:k-1}\Big(\mathrm{R}_k\big((I-A)^T(I-A)\big)\Big) = \mathrm{col}_{1:k-1}(G) - \mathrm{col}_{1:k-1}\Big(\mathrm{R}_k\big(A\big)\Big),$$

where $G = \mathrm{R}_k\big((\mathrm{R}_{k+1:p}(A))^T\big)\mathrm{R}_{k+1:p}\big(A\big)$. Note that $\mathrm{col}_{1:k-1}\Big(\big(R_{1:k-1}(G^T)\big)^T\Big) = \mathrm{col}_{1:k-1}(G)$. Therefore, we have

$$\mathrm{col}_{1:k-1}\Big(\mathrm{R}_k\big(A\big)\Big) = \mathrm{col}_{1:k-1}\Big(\big(R_{1:k-1}(G^T)\big)^T\Big) - \mathrm{col}_{1:k-1}\Big(\mathrm{R}_k\big((I-A)^T(I-A)\big)\Big).$$

By assumption, $\mathrm{R}_k\big((I-A)^T(I-A)\big)$ belongs to $\mathcal{L}_\alpha(M')$. We claim that $\big(R_{1:k-1}(G^T)\big)^T$ also belongs to $\mathcal{L}_\alpha(M')$. Combining above result with the fact that only the first $k-1$ columns of $\mathrm{R}_k(A)$ are non-zero, we have proven that $\mathrm{R}_k(A) \in \mathcal{L}_\alpha(M')$.

By induction, we finish the proof that $A \in \mathcal{L}_\alpha(M')$.

It remains to show that $\big(R_{1:k-1}(G^T)\big)^T \in \mathcal{L}_\alpha(M')$, where $G = \mathrm{R}_k\big(\mathrm{R}_{k+1:p}(A)^T\big)\mathrm{R}_{k+1:p}\big(A\big)$. To see this, assume $A \equiv [a_{ij}]_{p\times p}$ and $\big(R_{1:k-1}(G^T)\big)^T \equiv [s_{ij}]_{p\times p}$. Note that $s_{ij} = 0$ when $i \neq k$ or $j \geq k$. For the rest of k_{ij}, one can verify that

$$\sum_{j \leq (k-h)} |s_{kj}| \leq \sum_{i=1}^{\infty} M'^2 i^{-\alpha}(i+h)^{-\alpha} \leq \zeta(\alpha) M'^2 h^{-\alpha},$$

where $\alpha > 1$, $\zeta(\alpha)$ is the sum of hyperharmonic series with power of α. When $M' < \zeta(\alpha)^{-1}$, since the above inequality holds for any $0 < h < k$, It follows $\big(R_{1:k-1}(G^T)\big)^T \in \mathcal{L}_\alpha(M')$. $\square$

APPENDIX B

PRELIMINARY LEMMAS

In this section, we provide three preliminary lemmas and their proofs, which are important in the proofs of the other lemmas.

B.1 LEMMA 25

Lemma 23. *Operator $\mathbf{P}_\eta(\cdot)$ is defined in (2.3). For any square matrix $A \equiv [a_{ij}]_{p\times p}$ such that $\eta^{-1} \le \lambda_{\min}(A) \le \lambda_{\max}(A) \le \eta$, we have,*

$$\|A - \mathbf{P}_\eta(S)\|_{\mathrm{op}} \le 2\|A - S\|_{\mathrm{op}}, \tag{B.1}$$

$$\|A - \mathbf{P}_\eta(S)\|_{\mathrm{F}} \le 2\|A - S\|_{\mathrm{F}}. \tag{B.2}$$

Proof. Since

$$\|A - \mathbf{P}_\eta(S)\|_* \le ||S - \mathbf{P}_\eta(S)||_* + ||A - S||_*,$$

we only need to prove that

$$||S - \mathbf{P}_\eta(S)||_* \le ||A - S||_*,$$

where $||\cdot||_*$ is either the operator norm or Frobenius norm.

For an asymmetric matrix S, assume the singular value decomposition of S is $U^T SV = W$. Since operator norm and Frobenius norm are invariant to the orthogonal transformation, it is sufficient to prove

$$||U^T SV - U^T \mathbf{P}_\eta(S)V||_* \le ||U^T AV - U^T SV||_*.$$

Here, $U^T SV$ and $U^T \mathbf{P}_\eta(S)V$ are the diagonal matrices. Note that $\eta^{-1} \le \lambda_{\min}(U^T AV) \le \lambda_{\max}(U^T AV) \le \eta$. Without loss of generality, we only need to prove that for any diagonal matrix $W \equiv \mathrm{diag}(w)$, where $w = (w_1, \dots, w_p)^T$,

$$||W - W_\eta||_* \le ||A - W||_*,$$

where $W_\eta = \mathrm{diag}(\max\{\min\{w_i, \eta\}, \eta^{-1}\})$ and $\eta^{-1} \le \lambda_{\min}(A) \le \lambda_{\max}(A) \le \eta$. For the operator norm, note the fact that,

$$\begin{aligned}
\|A - W\|_{\mathrm{op}} &= \sup_{v \in \mathbb{R}^p, \|v\|_2 = 1} \|Av - Wv\|_2 \\
&\ge \max_i \|Ae_i - We_i\|_2 \\
&\ge \max_i \max\left\{\|Ae_i\|_2 - \|We_i\|_2, \|We_i\|_2 - \|Ae_i\|_2\right\} \\
&\ge \max_i \max\left\{\eta^{-1} - w_i, w_i - \eta, 0\right\} \\
&= \max_i \left\{|w_i - \max\{\min\{w_i, \eta\}, \eta^{-1}\}|\right\} \\
&= \|W_\eta - W\|_{\mathrm{op}}.
\end{aligned}$$

where e_i denotes the vector all of whose components are zero, except the i-th component being one. For the Frobenius norm, we have

$$\begin{aligned}
\|A - W\|_{\mathrm{F}}^2 &= \sum_{i \ne j} a_{ij}^2 + \sum_i (a_{ii} - w_i)^2 \\
&= \sum_i (\sum_j a_{ij}^2 + w_i^2 - 2a_{ii}w_i) \\
&\ge \sum_i (\sum_j a_{ij}^2 + w_i^2 - 2\sqrt{\sum_j a_{ij}^2} w_i) \\
&= \sum_i (w_i - \sqrt{\sum_j a_{ij}^2})^2.
\end{aligned}$$

Since $\eta^{-1} \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq \eta$, we obtain $\eta^{-1} \leq \sqrt{\sum_j a_{ij}^2} \leq \eta$. It is easy to check that for any i, $(\sqrt{\sum_j a_{ij}^2} - w_i)^2 \geq (\max\{\min\{w_i, \eta\}, \eta^{-1}\} - w_i)^2$. Then we have

$$\|A - W\|_{\mathrm{F}}^2 \geq \|W_\eta - W\|_{\mathrm{F}}^2,$$

that yields $\|A - W\|_{\mathrm{F}} \geq \|W_\eta - W\|_{\mathrm{F}}$.

The same result can be derived easily in the case of symmetric matrix S using similar argument on the eigen-decomposition. Thus, we finish the proof. □

B.2 LEMMA 26

Lemma 24. *Assume precision matrix Ω satisfies that $\eta^{-1} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq \eta$. Let $\Sigma = \Omega^{-1}$ and $\Omega = (I - A)^T D^{-1}(I - A)$ (see (1.1)), then,*

$$\eta^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \eta, \tag{B.3}$$

$$\eta^{-1} \leq \lambda_{\min}(D) \leq \lambda_{\max}(D) \leq \eta, \tag{B.4}$$

$$\eta^{-1} \leq \lambda_{\min}(I - A) \leq \lambda_{\max}(I - A) \leq \eta. \tag{B.5}$$

Proof. Since $\Sigma = \Omega^{-1}$, it is trivial that $\eta^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \eta$. For the symmetric positive-definite matrices $\Omega \equiv [\omega_{ij}]_{p\times p}$ and $\Sigma \equiv [\sigma_{ij}]_{p\times p}$, $\omega_{ii} \leq \lambda_{\max}(\Omega) \leq \eta; \sigma_{ii} \leq \lambda_{\max}(\Sigma) \leq \eta$ for any $i \in [p]$. ω_{ii}^{-1} can be interpreted as the variance of the residual of the regression of X_i on $\mathbf{X}_{-i} = (X_1 \dots X_{i-1}, X_{i+1} \dots X_p)^T$. σ_{ii} can be interpreted as the variance of the residual of the regression of X_i on $\mathbf{0}$. Therefore, the i-th entry of D, the variance of the residual of the regression of X_i on $\mathbf{X}_{1:i-1} = (X_1 \dots X_{i-1})^T$, has the bound $\omega_{ii}^{-1} \leq d_i \leq \sigma_{ii}$, then $\eta^{-1} \leq \lambda_{\min}(D) \leq \lambda_{\max}(D) \leq \eta$.

For any unit vector u, v such that $(I - A)u = \lambda v$, $\lambda > 0$, we have

$$u^T \Omega u = u^T (I - A)^T D^{-1}(I - A)u = \lambda^2 v^T D^{-1} v,$$

we obtain that $\lambda^2 = (u^T \Omega u)/(v^T D^{-1} v) \in [\eta^{-2}, \eta^2]$, that yields

$$\eta^{-1} \leq \lambda_{\min}(I - A) \leq \lambda_{\max}(I - A) \leq \eta.$$ □

B.3 LEMMA 27

Lemma 25. *Assume* $\mathbf{X}$ *is an* i*-variate random vector with covariance matrix* Σ *such that* $\eta^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \eta$. *Let the linear projection of* X_i *onto* $\mathbf{X}_{i-k:i-1}$ *in population be* $\hat{X}_i^{\langle k \rangle}$. *With the corresponding coefficients padded with* $i-k-1$ *zeros in the front, we can rewrite it as* $\hat{X}_i^{\langle k \rangle} = \mathbf{X}_{1:i-1}^T \boldsymbol{\beta}_i^{\langle k \rangle}$, *where* $\boldsymbol{\beta}_i^{\langle k \rangle} \in \mathbb{R}^{i-1}$ *with its first* $i-k-1$ *coordinates being zeros. In addition, set* $\epsilon_i^{\langle k \rangle} = \mathsf{Var}(X_i - \hat{X}_i^{\langle k \rangle})$.

(i) Whenever $\boldsymbol{\beta}_i^{\langle i-1 \rangle} = \left(\beta_{i1}^{\langle i-1 \rangle}, \ldots, \beta_{i(i-1)}^{\langle i-1 \rangle}\right)^T$ *satisfies*

$$\sum_{j<i-k} |\beta_{ij}^{\langle i-1 \rangle}| < Mk^{-\alpha}, \quad k \in [i-1], \tag{B.6}$$

we have,

$$\|\boldsymbol{\beta}_i^{\langle i-1 \rangle} - \boldsymbol{\beta}_i^{\langle k \rangle}\|_2 \leq 2\eta^2 Mk^{-\alpha}, \tag{B.7}$$

$$|\epsilon_i^{\langle i-1 \rangle} - \epsilon_i^{\langle k \rangle}| \leq 4\eta^4 Mk^{-\alpha}. \tag{B.8}$$

(ii) Whenever $\boldsymbol{\beta}_i^{\langle i-1 \rangle} = \left(\beta_{i1}^{\langle i-1 \rangle}, \ldots, \beta_{i(i-1)}^{\langle i-1 \rangle}\right)^T$ *satisfies*

$$|\beta_{ij}^{\langle i-1 \rangle}| < M(i-j)^{-\alpha-1}, \quad k \in [i-1], \tag{B.9}$$

it holds that,

$$\|\boldsymbol{\beta}_i^{\langle i-1 \rangle} - \boldsymbol{\beta}_i^{\langle k \rangle}\|_2 \leq 2\eta^2 M(k-1)^{-\alpha-\frac{1}{2}}, \tag{B.10}$$

$$|\epsilon_i^{\langle i-1 \rangle} - \epsilon_i^{\langle k \rangle}| \leq 4\eta^4 M(k-1)^{-\alpha-\frac{1}{2}}. \tag{B.11}$$

Proof. For any fixed i and k, let $\boldsymbol{\beta}_i^{\langle i-1 \rangle} = ((\boldsymbol{\beta}_A^{\langle i-1 \rangle})^T, (\boldsymbol{\beta}_B^{\langle i-1 \rangle})^T)^T$, $\boldsymbol{\beta}_i^{\langle k \rangle} = ((\boldsymbol{\beta}_A^{\langle k \rangle})^T, (\boldsymbol{\beta}_B^{\langle k \rangle})^T)^T$, where the sizes of $\boldsymbol{\beta}_A^{\langle i-1 \rangle}$ and $\boldsymbol{\beta}_A^{\langle k \rangle}$ are $i-k-1$, the sizes of $\boldsymbol{\beta}_B^{\langle i-1 \rangle}$ and $\boldsymbol{\beta}_B^{\langle k \rangle}$ are k. Note that $\boldsymbol{\beta}_A^{\langle k \rangle} = \mathbf{0}$. To facilitate the proof, we divide Σ into several block matrices,

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{bmatrix},$$

where $\Sigma_{11}, \Sigma_{22}, \Sigma_{33}$ are the covariance matrices of $\mathbf{X}_{1:i-k-1}, \mathbf{X}_{i-k:i-1}$ and X_i respectively.

By the definition of the linear projection, we have

$$\begin{aligned}
\Sigma_{23} &= \Sigma_{21}\boldsymbol{\beta}_A^{\langle i-1\rangle} + \Sigma_{22}\boldsymbol{\beta}_B^{\langle i-1\rangle}, \\
\epsilon^{\langle i-1\rangle} &= (-(\boldsymbol{\beta}_i^{\langle i-1\rangle})^T, 1)\Sigma(-(\boldsymbol{\beta}_i^{\langle i-1\rangle})^T, 1)^T, \\
\Sigma_{23} &= \Sigma_{22}\boldsymbol{\beta}_B^{\langle k\rangle}, \\
\epsilon^{\langle k\rangle} &= (-(\boldsymbol{\beta}_i^{\langle k\rangle})^T, 1)\Sigma(-(\boldsymbol{\beta}_i^{\langle k\rangle})^T, 1)^T.
\end{aligned}$$

Condition in (B.6) implies

$$\|\boldsymbol{\beta}_A^{\langle i-1\rangle}\|_2 \le \|\boldsymbol{\beta}_A^{\langle i-1\rangle}\|_1 \le Mk^{-\alpha}.$$

Then we have,

$$\begin{aligned}
\|\boldsymbol{\beta}_i^{\langle i-1\rangle} - \boldsymbol{\beta}_i^{\langle k\rangle}\|_2 &\le \|\boldsymbol{\beta}_A^{\langle i-1\rangle} - \mathbf{0}\|_2 + \|\boldsymbol{\beta}_B^{\langle i-1\rangle} - \boldsymbol{\beta}_B^{\langle k\rangle}\|_2 \\
&\le \|\boldsymbol{\beta}_A^{\langle i-1\rangle}\|_2 + \|\Sigma_{22}^{-1}\Sigma_{21}\boldsymbol{\beta}_A^{\langle i-1\rangle}\|_2 \\
&\le Mk^{-\alpha} + \eta^2 Mk^{-\alpha} \\
&\le 2\eta^2 Mk^{-\alpha}.
\end{aligned}$$

that yields the desired (B.7).

Assume the decomposition of Σ^{-1} is $(I-A)^T D^{-1}(I-A)$. Note that $(-(\boldsymbol{\beta}_i^{\langle i-1\rangle})^T, 1)$ corresponds to the ith row of the lower triangle of $I-A$. According to Lemma 23, $\|I-A\|_{\rm op} \le \eta$, then $\|(-(\boldsymbol{\beta}_i^{\langle i-1\rangle})^T, 1)^T\|_2 \le \|I-A\|_{\rm op} \le \eta$. Applying the same argument on the covariance matrix of $\mathbf{X}_{i-k:i}$, we have $\|(-(\boldsymbol{\beta}_i^{\langle k\rangle})^T, 1)^T\|_2 \le \eta$. Moreover,

$$\begin{aligned}
&|\epsilon_i^{\langle i-1\rangle} - \epsilon_i^{\langle k\rangle}| \\
\le& |(-(\boldsymbol{\beta}_i^{\langle i-1\rangle})^T, 1)\Sigma(-(\boldsymbol{\beta}_i^{\langle i-1\rangle})^T, 1)^T - (-(\boldsymbol{\beta}_i^{\langle k\rangle})^T, 1)\Sigma(-(\boldsymbol{\beta}_i^{\langle k\rangle})^T, 1)^T| \\
\le& \big(\|(-(\boldsymbol{\beta}_i^{\langle i-1\rangle})^T, 1)^T\|_2 + \|(-(\boldsymbol{\beta}_i^{\langle k\rangle})^T, 1)^T\|_2\big)\|\Sigma\|_{\rm op}\|\boldsymbol{\beta}_i^{\langle i-1\rangle} - \boldsymbol{\beta}_i^{\langle k\rangle}\|_2 \\
\le& 4\eta^4 Mk^{-\alpha},
\end{aligned}$$

that yields the desired (B.8).

Similarly, condition in (B.9) implies

$$\|\boldsymbol{\beta}_A^{\langle i-1\rangle}\|_2 \le \|\boldsymbol{\beta}_A^{\langle i-1\rangle}\|_1 \le M(k-1)^{-\alpha-\frac{1}{2}}.$$

Then we have

$$\|\boldsymbol{\beta}_i^{\langle i-1\rangle} - \boldsymbol{\beta}_i^{\langle k\rangle}\|_2 \leq 2\eta^2 M(k-1)^{-\alpha-\frac{1}{2}},$$
$$|\epsilon_i^{\langle i-1\rangle} - \epsilon_i^{\langle k\rangle}| \leq 4\eta^4 M(k-1)^{-\alpha-\frac{1}{2}},$$

which completes the proofs of (B.10) and (B.11). □

Proofs of Lemma 2 and Lemma 4 in analysis of Theorem 3 In this section, we prove Lemma 2 and Lemma 4 to establish Theorem 3.

APPENDIX C

LEMMAS FOR THEOREM 3

In this section, we prove Lemma 2 and Lemma 4 to establish Theorem 3.

C.1 PROOF OF LEMMA 2

For $\Omega \equiv [\omega_{ij}]_{p\times p}$, define

$$bd_k(\Omega) \equiv [\omega_{ij}\mathbf{1}(|i-j| \le k)]_{p\times p}. \tag{C.1}$$

It is easy to check $\Omega_k^* = \frac{1}{k}\sum_{i=k}^{2k-1} bd_i(\Omega)$. Then we have

$$\|\Omega - \Omega_k^*\|_{\mathrm{op}} \le \frac{1}{k}\sum_{i=k}^{2k-1}(\|\Omega - bd_i(\Omega)\|_{\mathrm{op}}).$$

We turn to the analysis of $\|\Omega - bd_k(\Omega)\|_{\mathrm{op}}$. Define $D^{-\frac{1}{2}}(I-A)$ as $B \equiv [b_{ij}]$. We know $\Omega = B^T B$, and

$$\max_i \sum_{j<i-k} |b_{ij}| \le M\eta^{\frac{1}{2}} k^{-\alpha}.$$

Set $bd_k(B)$ as B_k, and $B - bd_k(B)$ as B_{-k}. Then we can rewrite Ω as

$$\begin{aligned}\Omega &= B^T B \\ &= (B_k + B_{-k})^T(B_k + B_{-k}) \\ &= B_k^T B_k + B_{-k}^T B_k + B_k^T B_{-k} + B_{-k}^T B_{-k}.\end{aligned}$$

Checking the entries of $B_k^T B_k$, we find that $bd_k\big(B_k^T B_k\big) = B_k^T B_k$. Thus,

$$\begin{aligned}
&\|\Omega - bd_k(\Omega)\|_{\mathrm{op}} \qquad\qquad (C.2)\\
=&\|B^T B - bd_k(B^T B)\|_{\mathrm{op}}\\
=&\|B_{-k}^T B_k + B_k^T B_{-k} + B_{-k}^T B_{-k} - bd_k(B_{-k}^T B_k + B_k^T B_{-k} + B_{-k}^T B_{-k})\|_{\mathrm{op}}\\
\le&\|B_{-k}^T B_k + B_k^T B_{-k} + B_{-k}^T B_{-k}\|_{\mathrm{op}} + \|bd_k(B_{-k}^T B_k + B_k^T B_{-k} + B_{-k}^T B_{-k})\|_{\mathrm{op}}\\
=&\|B_{-k}^T B + B^T B_{-k} - B_{-k}^T B_{-k}\|_{\mathrm{op}} + \|bd_k(B_{-k}^T B_k + B_k^T B_{-k} + B_{-k}^T B_{-k})\|_{\mathrm{op}}\\
\le&\|B_{-k}^T B\|_{\mathrm{op}} + \|B^T B_{-k}\|_{\mathrm{op}} + \|B_{-k}^T B_{-k}\|_{\mathrm{op}}\\
&+ \|bd_k(B_{-k}^T B_k)\|_{\mathrm{op}} + \|bd_k(B_k^T B_{-k})\|_{\mathrm{op}} + \|bd_k(B_{-k}^T B_{-k})\|_{\mathrm{op}}.
\end{aligned}$$

The key is to control $\|B_{-k}\|_{\mathrm{op}}$. In addition, in order to get rid off the operator $bd_k(\cdot)$ when handling the term $\|bd_k(B_{-k}^T B_{-k})\|_{\mathrm{op}}$, we adopt a technique which requires controlling the operator norm of the matrix in which each entry is the absolute value of the entry in B_{-k}. To this end, we state the key lemma below.

Lemma 26. *Assume that matrix B_{-k} is defined above. We have*

$$\|B_{-k}^+\|_{\mathrm{op}} \le C k^{\frac{1}{2}-\alpha},$$

where X^+ is the matrix in which each entry is the magnitude of the corresponding entry in X.

Proof. Assume F_i is the matrix composed by the $(2^{i-1}k+1)$-th to $(2^i k)$-th sub-diagonals in matrix B^+, $i \in [\lceil \log_2(p/k)\rceil]$. Note that $B_{-k}^+ = \sum_{i=1}^{\lceil \log_2(p/k)\rceil}(F_i)$. For any $i \ge 1$, $\|F_i\|_\infty \le M\eta^{\frac{1}{2}}(2^{i-1}k)^{-\alpha}$. Since there are at most $2^{i-1}k$ entries in each column of F_i, $\|F_i\|_1 \le M\eta^{\frac{1}{2}}(2^{i-1}k)^{1-\alpha}$. The operator norm of F_i can be bounded by the two terms above,

$$\begin{aligned}
\|F_i\|_{\mathrm{op}} &\le (\|F_i\|_\infty \|F_i\|_1)^{\frac{1}{2}}\\
&\le M\eta^{\frac{1}{2}}(2^{i-1}k)^{\frac{1}{2}-\alpha}\\
&= M\eta^{\frac{1}{2}} k^{\frac{1}{2}-\alpha} \times (2^{\frac{1}{2}-\alpha})^{i-1}.
\end{aligned}$$

Consequently,

$$
\begin{aligned}
\|B^+_{-k}\|_{\text{op}} &\le \sum_{i=1}^{\lceil \log_2(p/k)\rceil} \|F_i\|_{\text{op}} \\
&\le \sum_{i=1}^{\lceil \log_2(p/k)\rceil} (M\eta^{\frac{1}{2}} k^{\frac{1}{2}-\alpha} \times (2^{\frac{1}{2}-\alpha})^{i-1}) \\
&= M\eta^{\frac{1}{2}} k^{\frac{1}{2}-\alpha} \times \sum_{i=1}^{\lceil \log_2(p/k)\rceil} (2^{\frac{1}{2}-\alpha})^{i-1} \\
&\le CM\eta^{\frac{1}{2}} k^{\frac{1}{2}-\alpha}.
\end{aligned}
$$

□

The bounds of the operator norms of many other matrices can be derived from the above result combining the following lemma.

Lemma 27. *Let X be a square matrix, then*

$$\|X\|_{\text{op}} \le \|X^+\|_{\text{op}},$$

and

$$\|bd_k(X)\|_{\text{op}} \le \|X^+\|_{\text{op}},$$

where X^+ is the matrix in which each entry is the magnitude of the corresponding entry in X.

Proof. Assume $X \equiv [x_{ij}]_{p\times p}$, let $\|X\|_{\text{op}} = u^T X v$, then

$$\|X\|_{\text{op}} = u^T X v = \sum_{i,j} u_i x_{ij} v_j \le \sum_{i,j} |u_i||x_{ij}||v_j| = (u^+)^T X^+ v^+ \le \|X^+\|_{\text{op}}.$$

Assume $bd_k(X) = [x^*_{ij}]_{p\times p}$, $\|bd_k(X)\|_{\text{op}} = u^T(bd_k(X))v$, then

$$\|bd_k(X)\|_{\text{op}} = u^T(bd_k(X))v = \sum_{i,j} u_i x^*_{ij} v_j \le \sum_{i,j} |u_i||x_{ij}||v_j| \le \|X^+\|_{\text{op}}.$$

□

It follows from the previous two lemmas that $\|B_{-k}\|_{\text{op}} \leq \|B^{+}_{-k}\|_{\text{op}} \leq Ck^{\frac{1}{2}-\alpha}$. Then we have

$$\begin{aligned} \|B^T B_{-k}\|_{\text{op}} &\leq \|B\|_{\text{op}}\|B_{-k}\|_{\text{op}} \leq Ck^{\frac{1}{2}-\alpha}, \\ \|B^T_{-k} B\|_{\text{op}} &\leq \|B\|_{\text{op}}\|B_{-k}\|_{\text{op}} \leq Ck^{\frac{1}{2}-\alpha}, \\ \|B^T_{-k} B_{-k}\|_{\text{op}} &\leq \|B_{-k}\|_{\text{op}}\|B_{-k}\|_{\text{op}} \leq Ck^{1-2\alpha}. \end{aligned} \tag{C.3}$$

In addition, Lemma 27 implies that,

$$\|bd_k(B^T_{-k}B_{-k})\|_{\text{op}} \leq \|(B^T_{-k}B_{-k})^{+}\|_{\text{op}} \leq \|(B^{+}_{-k})^T B^{+}_{-k}\|_{\text{op}} \leq Ck^{1-2\alpha}. \tag{C.4}$$

Then we turn to bound $\|bd_k(B^T_{-k}B_k)\|_{\text{op}} + \|bd_k(B^T_k B_{-k})\|_{\text{op}}$. We control $\|bd_k(B^T_{-k}B_k)\|_{\infty}$ and $\|bd_k(B^T_{-k}B_k)\|_1$ first. For $h \in [p]$, we have

$$\begin{aligned} \|bd_k(B^T_{-k}B_k)\|_{\infty} \leq & \max_h \|\text{row}_h(bd_k(B^T_{-k}B))\|_1 \\ \leq & \sum_{i=h+1}^{h+k} (\sum_{j=h+k+1}^{h+2k} |b_{jh}||b_{ij}|) \\ = & \sum_{j=h+k+1}^{h+2k} (\sum_{i=h+1}^{h+k} |b_{jh}||b_{ij}|) \\ \leq & \sum_{j=h+k+1}^{h+2k} M^2\eta((j-h)^{-\alpha}(j-h-k)^{-\alpha}) \\ = & \sum_{j=1}^{k} M^2\eta((j+k)^{-\alpha}j^{-\alpha}) \\ \leq & M^2\eta k^{-\alpha}\sum_{j=1}^{k} j^{-\alpha} \\ \leq & Ck^{1-2\alpha}, \end{aligned}$$

and

$$
\begin{aligned}
\|bd_k(B_{-k}^T B_k)\|_1 &\le \max_h \|\mathrm{col}_h(bd_k(B_{-k}^T B))\|_1 \\
&\le \sum_{i=h-k}^{h-1} (\sum_{j=h+1}^{h+k} |b_{ji}||b_{jh}|) \\
&\le \sum_{j=h+1}^{h+k} (\sum_{i=h-k}^{h-1} |b_{ji}||b_{jh}|) \\
&\le \sum_{j=h+1}^{h+k} M^2 \eta (k^{-\alpha}(j-h)^{-\alpha}) \\
&\le \sum_{j=1}^{k} M^2 \eta (k^{-\alpha} j^{-\alpha}) \\
&\le C k^{1-2\alpha}.
\end{aligned}
$$

Therefore, we obtain that

$$
\begin{aligned}
\|bd_k(B_{-k}^T B_k)\|_{\mathrm{op}} &\le (\|bd_k(B_{-k}^T B_k)\|_\infty \|bd_k(B_{-k}^T B_k)\|_1)^{1/2} \le C k^{1-2\alpha}, \\
\|bd_k(B_k^T B_{-k})\|_{\mathrm{op}} &\le (\|bd_k(B_k^T B_{-k})\|_\infty \|bd_k(B_k^T B_{-k})\|_1)^{1/2} \le C k^{1-2\alpha}.
\end{aligned} \tag{C.5}
$$

Combining (C.2), (C.3), (C.4) and (C.5), we prove that

$$
\|\Omega - bd_k(\Omega)\|_{\mathrm{op}} \le C k^{\frac{1}{2}-\alpha},
$$

which follows $\|\Omega - \Omega_k^*\|_{\mathrm{op}}^2 \le C k^{1-2\alpha}$.

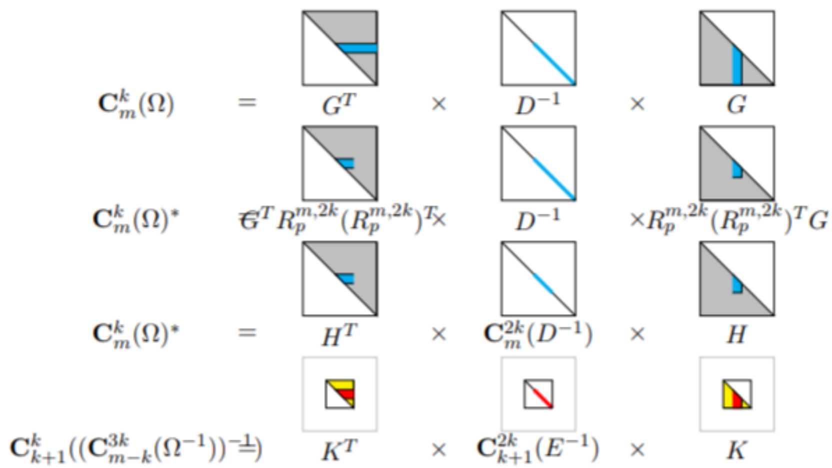

Figure 4: An illustration of the proof strategy in Lemma 4.

C.2 PROOF OF LEMMA 4

The proof strategy in this lemma is not complicated, although the notation is quite involved. Our target is to bound the distance between $\mathbf{C}_m^k(\Omega)$ and $\mathbf{C}_{k+1}^k\big((\mathbf{C}_{m-k}^{3k}(\Omega^{-1}))^{-1}\big)$ under the operator norm. To this end, we introduce an intermediate term $\mathbf{C}_m^k(\Omega)^*$ to facilitate our proof. Specifically, we break the target into two terms $\|\mathbf{C}_m^k(\Omega) - \mathbf{C}_m^k(\Omega)^*\|_{\mathrm{op}}^2$ and $\|\mathbf{C}_m^k(\Omega)^* - \mathbf{C}_{k+1}^k\big((\mathbf{C}_{m-k}^{3k}(\Omega^{-1}))^{-1}\big)\|_{\mathrm{op}}^2$ and derive their bounds respectively. The construction of $\mathbf{C}_m^k(\Omega)^*$ with corresponding Cholesky decomposition are illustrated in Figure 4, in contrast with those of $\mathbf{C}_m^k(\Omega)$ and $\mathbf{C}_{k+1}^k\big((\mathbf{C}_{m-k}^{3k}(\Omega^{-1}))^{-1}\big)$.

To express the decomposition of $\mathbf{C}_m^k(\Omega)$ in the matrix form, we define the $p \times k$ matrix

$$R_p^{m,k} \equiv [r_{ij}]_{p\times k}, \quad r_{ij} = \mathbf{1}\,(i - m = j - 1)\,.$$

Assume $\Omega = (I - A)^T D^{-1}(I - A)$. Set $(I - A)R_p^{m,k}$ as G. One can check

$$\begin{aligned}\mathbf{C}_m^k(\Omega) &= (R_p^{m,k})^T \Omega R_p^{m,k} \\ &= (R_p^{m,k})^T (I - A)^T D^{-1}(I - A) R_p^{m,k} \\ &= G^T D^{-1} G.\end{aligned}$$

Define

$$\begin{aligned}\mathbf{C}_m^k(\Omega)^* &= (R_p^{m,k})^T(I-A)^T R_p^{m,2k}(R_p^{m,2k})^T D^{-1} R_p^{m,2k}(R_p^{m,2k})^T(I-A)R_p^{m,k} \\ &= G^T R_p^{m,2k}(R_p^{m,2k})^T D^{-1} R_p^{m,2k}(R_p^{m,2k})^T G.\end{aligned}$$

We first bound $\|\mathbf{C}_m^k(\Omega)-\mathbf{C}_m^k(\Omega)^*\|_{\text{op}}^2$. Since $I-A$ is a lower triangular matrix, we know $R_p^{m,2k}(R_p^{m,2k})^T G$ consists of the first $2k$ columns of G. Then we have

$$G - R_p^{m,2k}(R_p^{m,2k})^T G = (0,0,\ldots,0,g_{m+2k},\ldots,g_p)^T,$$

where $g_i = \text{row}_i(G)$, and for $i \in (m+2k):p$,

$$\|g_i\|_2 \le \|g_i\|_1 \le M(i-m-k+1)^{-\alpha}. \tag{C.6}$$

Consequently,

$$\begin{aligned}&\|G - R_p^{m,2k}(R_p^{m,2k})^T G\|_{\text{op}}^2 \\ \le&\|(\|g_{m+2k}\|_2,\ldots,\|g_p\|_2)^T\|_2^2 \\ \le&\|(M(k+1)^{-\alpha},\ldots,M(p-m-k)^{-\alpha})^T\|_2^2 \\ \le&M^2 k^{-2\alpha+1}.\end{aligned}$$

Then we have

$$\begin{aligned}&\|\mathbf{C}_m^k(\Omega)-\mathbf{C}_m^k(\Omega)^*\|_{\text{op}}^2 \\ \le&\|(G^T D^{-1} G - G^T R_p^{m,2k}(R_p^{m,2k})^T D^{-1} R_p^{m,2k}(R_p^{m,2k})^T G\|_{\text{op}}^2 \\ \le&\|G - R_p^{m,2k}(R_p^{m,2k})^T G\|_{\text{op}}^2 \|D^{-1}\|_{\text{op}}^2(\|G\|_{\text{op}}^2 + \|R_p^{m,2k}(R_p^{m,2k})^T G\|_{\text{op}}^2) \\ \le&2\|D^{-1}\|_{\text{op}}^2\|G\|_{\text{op}}^2\|G - R_p^{m,2k}(R_p^{m,2k})^T G\|_{\text{op}}^2 \\ \le&2\eta^2 M^2 k^{-2\alpha+1}.\end{aligned} \tag{C.7}$$

Next, we turn to derive the bound of $\|\mathbf{C}_m^k(\Omega)^* - \mathbf{C}_{k+1}^k\big((\mathbf{C}_{m-k}^{3k}(\Omega^{-1}))^{-1}\big)\|_{\text{op}}^2$. Assume that $(\mathbf{C}_{m-k}^{3k}(\Omega^{-1}))^{-1} = (I-B)^T E^{-1}(I-B)$. One can also check

$$
\begin{aligned}
&\mathbf{C}_{k+1}^k\big((\mathbf{C}_{m-k}^{3k}(\Omega^{-1}))^{-1}\big)\\
=&(R_{3k}^{k+1,k})^T(\mathbf{C}_{m-k}^{3k}(\Omega^{-1}))^{-1}R_{3k}^{k+1,k}\\
=&(R_{3k}^{k+1,k})^T(I-B)^T E^{-1}(I-B)R_{3k}^{k+1,k}\\
=&(R_{3k}^{k+1,k})^T(I-B)^T R_{3k}^{k+1,2k}(R_{3k}^{k+1,2k})^T E^{-1} R_{3k}^{k+1,2k}(R_{3k}^{k+1,2k})^T(I-B)R_{3k}^{k+1,k}.
\end{aligned}
$$

To ease our notation, one can check $(R_p^{m,2k})^T D R_p^{m,2k} = \mathbf{C}_m^{2k}(D)$ and $(R_{3k}^{k+1,2k})^T E R_{3k}^{k+1,2k} = \mathbf{C}_{k+1}^{2k}(E)$. In addition, we set $(R_p^{m,2k})^T(I-A)R_p^{m,k}$ as H, $(R_{3k}^{k+1,2k})^T(I-B)R_{3k}^{k+1,k}$ as K. Then we can rewrite $\mathbf{C}_m^k(\Omega)^* = H^T\mathbf{C}_m^{2k}(D^{-1})H$ and $\|\mathbf{C}_m^k(\Omega) - \mathbf{C}_m^k(\Omega)^*\|_{\text{op}}^2 = K^T\mathbf{C}_{k+1}^{2k}(E^{-1})K$. We bound $\|H-K\|_{\text{op}}^2$ and $\|\mathbf{C}_m^{2k}(D^{-1}) - \mathbf{C}_{k+1}^{2k}(E^{-1})\|_{\text{op}}^2$ separately below.

Referring to the instruction in Figure 4, one can check $\text{row}_i(H)$ is part of the coefficients of the regression $X_{m+i} \sim \mathbf{X}_{1:m+i-1}$ and $\text{row}_i(K)$ is part of the coefficients of the regression $X_{m+i} \sim \mathbf{X}_{m-k:m+i-1}$. According to Lemma 25, $\text{row}_i(H-K)$ can be bounded by $2M\eta^2(k+i)^{-\alpha}$. The dimension of the non-zero part of $H-K$ is $2k\times k$. So we have

$$
\|H-K\|_{\text{op}}^2 \le 8\eta^4 M^2 k^{-2\alpha+1}.
$$

Similarly, the i-th elements of $\mathbf{C}_m^{2k}(D)$ and $\mathbf{C}_{k+1}^{2k}(E)$ are the residuals of the above two regressions. According to Lemma 25, $\|\mathbf{C}_m^{2k}(D) - \mathbf{C}_{k+1}^{2k}(E)\|_{\text{op}}^2 \le 16\eta^8 M^2 k^{-2\alpha}$. Therefore, we have

$$
\begin{aligned}
&\|\mathbf{C}_m^{2k}(D^{-1}) - \mathbf{C}_{k+1}^{2k}(E^{-1})\|_{\text{op}}^2\\
\le&\|\mathbf{C}_m^{2k}(D^{-1})\|_{\text{op}}^2\|\mathbf{C}_m^{2k}(D) - \mathbf{C}_{k+1}^{2k}(E)\|_{\text{op}}^2\|\mathbf{C}_{k+1}^{2k}(E^{-1})\|_{\text{op}}^2\\
\le&16M^2\eta^{12}k^{-2\alpha}.
\end{aligned}
$$

Combing the above two results, we have

$$
\begin{aligned}
&\|\mathbf{C}_m^k(\Omega)^* - \mathbf{C}_{k+1}^k\big((\mathbf{C}_{m-k}^{3k}(\Omega^{-1}))^{-1}\big)\|_{\text{op}}^2 \\
\leq &\|H^T\mathbf{C}_m^{2k}(D^{-1})H - K^T\mathbf{C}_{k+1}^{2k}(E^{-1})K\|_{\text{op}}^2 \\
\leq &\|H^T\|_{\text{op}}^2\|\mathbf{C}_m^{2k}(D^{-1}) - \mathbf{C}_{k+1}^{2k}(E^{-1})\|_{\text{op}}^2\|H\|_{\text{op}}^2 \\
&+ (\|H\|_{\text{op}}^2 + \|K\|_{\text{op}}^2)\|\mathbf{C}_{k+1}^{2k}(E^{-1})\|_{\text{op}}^2\|H-K\|_{\text{op}}^2 \\
\leq &\eta^4 \times 16M^2\eta^{12}k^{-2\alpha} + 6\eta^4 \times 8\eta^4M^2k^{-2\alpha+1} \\
\leq &96M^2\eta^{16}k^{-2\alpha+1}.
\end{aligned} \tag{C.8}
$$

In the end, based on the Equations (C.7) and (C.8), we have

$$
\begin{aligned}
&\|\mathbf{C}_m^k(\Omega) - \mathbf{C}_{k+1}^k\big((\mathbf{C}_{m-k}^{3k}(\Omega^{-1}))^{-1}\big)\|_{\text{op}}^2 \\
\leq &2\|\mathbf{C}_m^k(\Omega) - \mathbf{C}_m^k(\Omega)^*\|_{\text{op}}^2 + 2\|\mathbf{C}_m^k(\Omega)^* - \mathbf{C}_{k+1}^k\big((\mathbf{C}_{m-k}^{3k}(\Omega^{-1}))^{-1}\big)\|_{\text{op}}^2 \\
\leq &2 \times (2\eta^2M^2k^{-2\alpha+1} + 96M^2\eta^{16}k^{-2\alpha+1}) \\
\leq &200M^2\eta^{16}k^{-2\alpha+1}.
\end{aligned}
$$

We finish the proof of Lemma 4.

APPENDIX D

PROOFS FOR THEOREM 4

In this section, we prove Lemma 5, Lemma 7, Lemma 8, and Lemma 10 to establish Theorem 4.

D.1 PROOF OF LEMMA 5

First we prove that $\mathcal{P}_1 \in \mathcal{P}_\alpha(\eta, M)$. Let $A_k^*(\theta) \equiv [a_{ij}]_{k\times k}$. We know the exact value of each entry. It is easy to check $\sum_{i-j>k} |a_{ij}| \leq Mk^{-\alpha}$. One can check $\Omega(\theta) \in \mathcal{P}_1$ has the specific form of

$$\Omega(\theta) = \begin{bmatrix} I_k + (A_k^*(\theta))^T A_k^*(\theta) & -(A_k^*(\theta))^T & \mathbf{0}_{k\times(p-2k)} \\ -A_k^*(\theta) & I_k & \mathbf{0}_{k\times(p-2k)} \\ \mathbf{0}_{(p-2k)\times k} & \mathbf{0}_{(p-2k)\times k} & I_{p-2k} \end{bmatrix}.$$

Let $\mathbf{1}$ denote the vector with all 1's in Θ. One can check that

$$\begin{aligned}\lambda_{\max}(\Omega(\theta)) =& \big(\lambda_{\max}(I - A(\theta))\big)^2 \leq \big(\lambda_{\max}(I + A(\theta))\big)^2 \leq (1 + \lambda_{\max}(A_k^*(\mathbf{1})))^2 \\ =& (1 + k^{\frac{1}{2}} n^{-\frac{1}{2}} \tau)^2 \leq \eta.\end{aligned}$$

The second inequality above is due to that the entries of $A(\theta)$ are all non-negative and Lemma 27. Recall $\Sigma(\theta) = (I + A(\theta))(I + A(\theta))^T$. Thus, we can check

$$\begin{aligned}\lambda_{\min}(\Omega(\theta)) =& \big(\lambda_{\max}(\Sigma(\theta))\big)^{-1} = \big(\lambda_{\max}(I + A(\theta))\big)^{-2} \geq (1 + \lambda_{\max}(A_k^*(\mathbf{1})))^{-2} \\ =& (1 + k^{\frac{1}{2}} n^{-\frac{1}{2}} \tau)^{-2} \geq \eta^{-1}.\end{aligned}$$

The eigenvalues of $\Omega(\theta)$ are in the interval $[\eta^{-1}, \eta]$. So $\mathcal{P}_1 \in \mathcal{P}_\alpha(\eta, M)$.

Then we turn to prove that $\mathcal{P}_2 \in \mathcal{P}_\alpha(\eta, M)$. The Cholesky factor A of $\Omega(m)$ is the zero matrix. The minimum eigenvalue of $\Omega(m)$ is $(1 + \tau a^{\frac{1}{2}})^{-1}$, which is greater than η^{-1} and maximum one is 1, which is less than η. So $\mathcal{P}_2 \in \mathcal{P}_\alpha(\eta, M)$.

D.2 PROOF OF LEMMA 7

Since $\|P_\theta \wedge P_{\theta'}\| = 1 - \frac{1}{2}\|P_\theta - P_{\theta'}\|_1$, we turn to control $\max_{H(\theta,\theta')=1} \|P_\theta - P_{\theta'}\|_1$. The following First Pinsker's inequality will facilitate our analysis.

Lemma 28 (First Pinsker's Inequality [Csiszár, 1967])**.**

$$\begin{aligned}
\|P_\theta - P_{\theta'}\|_1^2 &\leq \frac{1}{2} K(P_{\theta'}|P_\theta) \\
&= \frac{n}{2}[\frac{1}{2} tr(\Sigma(\theta')\Sigma(\theta)^{-1}) - \frac{1}{2}\log\det(\Sigma(\theta')\Sigma(\theta)^{-1}) - \frac{p}{2}]
\end{aligned}$$

where $K(\cdot|\cdot)$ is the Kullback-Leibler divergence.

One can check $\Sigma(\theta) \in \mathcal{P}_1$ has the form of

$$\begin{aligned}
\Sigma(\theta) &= (I + A(\theta))(I + A(\theta)^T) \\
&= \begin{bmatrix} I_k & (A_k^*(\theta))^T & 0_{k\times(p-2k)} \\ (A_k^*(\theta)) & I_k + (A_k^*(\theta))(A_k^*(\theta))^T & 0_{k\times(p-2k)} \\ 0_{(p-2k)\times k} & 0_{(p-2k)\times k} & I_{p-2k}. \end{bmatrix}.
\end{aligned}$$

Set $\Sigma(\theta') = D + \Sigma(\theta)$ and $d = \theta' - \theta$. Note that whenever $\max H(\theta, \theta') = 1$, D has the form of

$$\begin{aligned}
D &= (A(\theta') - A(\theta)) + (A(\theta') - A(\theta))^T + (A(\theta')A(\theta')^T - A(\theta)A(\theta)^T) \\
&= A(d) + A(d)^T + A(\theta')A(d)^T + A(d)A(\theta)^T,
\end{aligned}$$

where $A(d)$ is similarly defined as $A(\theta)$ except that θ_i is replaced by d_i. Let $\mathbf{1}$ denote the all-ones vector, one can check

$$\|D\|_{\mathrm{op}} \leq 2(1+\|A_k^*(\mathbf{1})\|_{\mathrm{op}})\|A_k^*(d)\|_{\mathrm{op}} \leq 4\tau n^{-\frac{1}{2}}, \tag{D.1}$$

$$\|D\|_{\mathrm{F}} \leq 2(1+\|A_k^*(\mathbf{1})\|_{\mathrm{F}})\|A_k^*(d)\|_{\mathrm{F}} \leq 4\tau n^{-\frac{1}{2}}. \tag{D.2}$$

Furthermore, we have $\|D\Sigma(\theta)^{-1}\|_{\mathrm{op}} \leq \eta 4\tau n^{-\frac{1}{2}} \leq n^{-\frac{1}{2}}$, $\|D\Sigma(\theta)^{-1}\|_{\mathrm{F}} \leq \|\Sigma(\theta)^{-1}\|_{\mathrm{op}}\|D\|_{\mathrm{F}} \leq \eta \times 4\tau n^{-\frac{1}{2}} \leq n^{-\frac{1}{2}}$. One can easily check that $\Sigma(\theta')\Sigma(\theta)^{-1} = I + D\Sigma(\theta)^{-1}$, $\frac{1}{2}tr(I+D\Sigma(\theta)^{-1}) = \frac{p}{2} + \frac{1}{2}tr(D\Sigma(\theta)^{-1})$, $\log\det(I + D\Sigma(\theta)^{-1}) = tr(D\Sigma(\theta)^{-1}) + \sum_i(\log(1+\lambda_i) - \lambda_i)$, where λ_i's are the eigenvalues of $D\Sigma(\theta)^{-1}$. Applying the First Pinsker's inequality in this case, we have

$$\begin{aligned}
\|P_\theta - P_{\theta'}\|_1^2 &\leq \frac{n}{2}[\frac{1}{2}tr(\Sigma(\theta')\Sigma(\theta)^{-1}) - \frac{1}{2}\log\det(\Sigma(\theta')\Sigma(\theta)^{-1}) - \frac{p}{2}] \\
&= \frac{n}{2}[\frac{1}{2}tr(I + D\Sigma(\theta)^{-1}) - \frac{1}{2}\log\det(I + D\Sigma(\theta)^{-1}) - \frac{p}{2}] \\
&= \frac{n}{2}\sum_i(\lambda_i - \log(1+\lambda_i)).
\end{aligned}$$

Since $|\lambda_i| \leq \|D\Sigma(\theta)^{-1}\|_{\mathrm{op}}$, all the λ_i are bounded by $\pm n^{-\frac{1}{2}}$. By Taylor expansion, $\sum_i(\lambda_i - \log(1+\lambda_i)) \leq 2\sum_i \lambda_i^2 = 2\|D\Sigma(\theta)^{-1}\|_{\mathrm{F}}^2$, we have

$$\|P_\theta - P_{\theta'}\|_1^2 \leq n\|D\Sigma(\theta)^{-1}\|_{\mathrm{F}}^2 \leq 1,$$

which implies $\|P_\theta \wedge P_{\theta'}\| \geq 0.5$.

D.3 PROOF OF LEMMA 8

The proof is as follows,

$$\begin{aligned}
&\|\Omega(\theta') - \Omega(\theta)\|_{\mathrm{op}} \\
\geq & \sup_{\|u\|_2=1, \|v\|_2=1} (u^T, \mathbf{0}^T, \mathbf{0}^T)(\Omega(\theta') - \Omega(\theta))(\mathbf{0}, v, \mathbf{0})^T \\
\geq & \sup_{\|u\|_2=1, \|v\|_2=1} u^T (A_k^*(\theta) - A_k^*(\theta'))^T v \\
= & \|(A_k^*(\theta') - A_k^*(\theta))\|_{\mathrm{op}} \\
= & (H(\theta', \theta))^{1/2} \tau n^{-\frac{1}{2}},
\end{aligned}$$

which immediately implies,

$$\min_{H(\theta,\theta')\geq 1} \frac{\|\Omega(\theta) - \Omega(\theta')\|_2^2}{H(\theta, \theta')} \geq (\tau n^{-\frac{1}{2}})^2.$$

D.4 PROOF OF LEMMA 10

Denote the density functions of P_i and $\bar{P}$ by f_i and $\bar{f}$ respectively, where $0 \leq i \leq p$. It is sufficient to bound $\int \frac{\bar{f}^2}{f_0} du - 1$ because $\|P_0 \wedge \bar{P}\| \geq 1 - \frac{1}{2}(\int \frac{\bar{f}^2}{f_0} du - 1)^{\frac{1}{2}}$. To this end, note that

$$\begin{aligned}
\int \frac{\bar{f}^2}{f_0} du - 1 &= \frac{1}{p^2}(\sum_{k\in[p]} \int \frac{f_k^2}{f_0} du + \sum_{1\leq i\neq j\leq p} \int \frac{f_i f_j}{f_0} du) - 1 \\
&= \frac{1}{p^2}(p(1 - \tau^2 k n^{-1})^{-\frac{n}{2}} + p^2 - p) - 1 \\
&\leq \frac{1}{2p}\tau^2 k \leq \frac{1}{16},
\end{aligned}$$

which further implies $\|P_0 \wedge \bar{P}\| \geq 1 - \frac{1}{2} \times \frac{1}{4} = \frac{7}{8}$.

APPENDIX E

PROOFS FOR THEOREM 5

In this section, we prove Lemma 11 and Lemma 12 to establish Theorem 5.

E.1 PROOF OF LEMMA 11

Recall $\Omega_k^* = \frac{1}{k}\sum_{h=k}^{2k-1} bd_h(\Omega)$, where $bd_h(\cdot)$ is defined in Equation (C.1). Then,

$$\|\Omega_k^* - \Omega\|_{\mathrm{op}} \le \frac{1}{k}\sum_{h=k}^{2k-1} \|bd_h(\Omega) - \Omega\|_{\mathrm{op}}.$$

For any fix $k \le h < 2k$, define $U_h \equiv [\omega_{ij}\mathbf{1}\,(i-j>h)]_{p\times p}$, then $bd_h(\Omega) - \Omega = U_h + U_h^T$. Note that $\Omega = (I-A)^T D^{-1}(I-A)$, where $I - A \equiv [a'_{ij}]_{p\times p}$. We have

$$\begin{aligned}
\|U_h\|_\infty &\le \max_i \|\mathrm{row}_i(U_h)\|_1 \le \max_i \eta \sum_{s=i}^{p} |a'_{si}| \sum_{j=1}^{i-h} |a'_{sj}| \\
&\le \max_i \eta M \sum_{s=i}^{p} (s-i)^{-\alpha-1}(s-i+h)^{-\alpha} \\
&\le C\eta M h^{-\alpha}. && \text{(E.1)}
\end{aligned}$$

Similarly, we have

$$
\begin{aligned}
\|U_h\|_1 &\le \max_j \|\mathrm{col}_j(U_h)\|_1 \le \max_j \eta M \sum_{s=j+h}^{p} |a'_{sj}| \\
&\le \max_j \eta M \sum_{s=j+h}^{p} (s-j)^{-\alpha-1} \\
&\le C\eta M h^{-\alpha}. \qquad \text{(E.2)}
\end{aligned}
$$

With the bounds for $\|U_h\|_1$ and $\|U_h\|_\infty$, we have

$$
\begin{aligned}
\|\Omega_k^* - \Omega\|_{\mathrm{op}} &\le \frac{1}{k}\sum_{i=k}^{2k-1} \|bd_i(\Omega) - \Omega\|_{\mathrm{op}} \\
&\le \max_{k\le h<2k} \|bd_h(\Omega) - \Omega\|_{\mathrm{op}} \\
&\le \max_{k\le h<2k} (\|U_h\|_{\mathrm{op}} + \|U_h^T\|_{\mathrm{op}}) \\
&\le \max_{k\le h<2k} 2\sqrt{\|U_h\|_\infty \|U_h\|_1} \\
&\le C\eta M k^{-\alpha}.
\end{aligned}
$$

E.2 PROOF OF LEMMA 12

The proof of Lemma 12 is basically the same as the one of Lemma 4, except for a few steps, which are highlighted below.

Since $\Omega \in \mathcal{Q}_\alpha(\eta, M)$, (C.6) should be updated by

$$\|g_i\|_2 \le M(i-m-k+1)^{-\alpha-1/2},$$

and consequently (C.7) should be replaced by

$$\|\mathbf{C}_m^k(\Omega) - \mathbf{C}_m^k(\Omega)^*\|_{\mathrm{op}}^2 \le 2M^2 k^{-2\alpha}.$$

According to Lemma 25, (C.8) is replaced by

$$\|\mathbf{C}_m^k(\Omega)^* - \mathbf{C}_{k+1}^k((\mathbf{C}_{m-k}^{3k}(\Omega^{-1}))^{-1})\|_{\mathrm{op}}^2 \le 96M^2\eta^{16}k^{-2\alpha}.$$

Therefore, we have

$$\|\mathbf{C}_m^k(\Omega) - \mathbf{C}_{k+1}^k((\mathbf{C}_{m-k}^{3k}(\Omega^{-1}))^{-1})\|_{\mathrm{op}}^2 \le 200M^2\eta^{16}k^{-2\alpha}.$$

APPENDIX F

PROOFS FOR THEOREM 6

In this section, we prove Lemma 13 and Lemma 14 to establish Theorem 6.

F.1 PROOF OF LEMMA 13

Let $B_k^*(\theta) \equiv [b_{ij}]_{k\times k}$, then we have the specific value of each entry. It is easy to check $b_{ij} \le M(i-j)^{-\alpha-1}$. One can check $\Omega(\theta) \in \mathcal{P}_3$ has the specific form of

$$\Omega(\theta) = \begin{bmatrix} I_k + B_k^*(\theta)^T B_k^*(\theta) & -B_k^*(\theta)^T & 0_{k\times(p-2k)} \\ -B_k^*(\theta) & I_k & 0_{k\times(p-2k)} \\ 0_{(p-2k)\times k} & 0_{(p-2k)\times k} & I_{p-2k} \end{bmatrix}. \tag{F.1}$$

Let $\mathbf{1}$ denote the vector with all 1's in Θ. Then one can check that

$$\begin{aligned}\lambda_{\max}(\Omega(\theta)) =& \big(\lambda_{\max}(I - B(\theta))\big)^2 \le \big(\lambda_{\max}(I + B(\theta))\big)^2 \le (1 + \lambda_{\max}(B_k^*(\mathbf{1})))^2 \\ =& (1 + k^{\frac{1}{2}} n^{-\frac{1}{2}}\tau)^2 \le \eta.\end{aligned}$$

The second inequality above follows from that the entries of $A(\theta)$ are all non-negative and Lemma 27. Recall $\Sigma(\theta) = (I + B(\theta))(I + B(\theta))^T$. Therefore, we have

$$\begin{aligned}\lambda_{\min}(\Omega(\theta)) =& \big(\lambda_{\max}(\Sigma(\theta))\big)^{-1} = \big(\lambda_{\max}(I + B(\theta))\big)^{-2} \ge (1 + \lambda_{\max}(B_k^*(\mathbf{1})))^{-2} \\ =& (1 + k^{\frac{1}{2}} n^{-\frac{1}{2}}\tau)^{-2} \ge \eta^{-1}.\end{aligned}$$

The eigenvalues of $\Omega(\theta)$ are in the interval $[\eta^{-1}, \eta]$. Therefore, $\mathcal{P}_3 \in \mathcal{P}_\alpha(\eta, M)$.

F.2 PROOF OF LEMMA 14

The proof follows most part of the one of Lemma 7. Some inequalities in the proof of Lemma 7 need to be rechecked. Assume that $\Sigma(\theta) = \{\Omega(\theta)^{-1} : \Omega(\theta) \in \mathcal{P}_3\}$ and $D = \Sigma(\theta') - \Sigma(\theta)$, one can verify that D has the same decomposition as that in the proof of Lemma 7 except that A_k^* is replaced by B_k^*. We can show

$$\|D\Sigma(\theta)^{-1}\|_{\text{op}} \leq \eta\|D\|_\infty \leq \eta(\tau(nk)^{-\frac{1}{2}}k + (\tau(nk)^{-\frac{1}{2}})^2k^2) \leq \frac{1}{2},$$

$$\|D\Sigma(\theta)^{-1}\|_{\text{F}}^2 \leq \eta^2(2k(\tau(nk)^{-\frac{1}{2}})^2 + (2k-1)((\tau(nk)^{-\frac{1}{2}})^2k)^2) \leq 4\eta^2(\tau n^{-\frac{1}{2}})^2.$$

Thus, (2.22) still holds in this case. As for (2.23), similarly,

$$\|\Omega(\theta') - \Omega(\theta)\|_{\text{op}}^2 \geq \|B_k^*(\theta') - B_k^*(\theta)\|_{\text{op}}^2 = H(\theta,\theta')(\tau(nk)^{-\frac{1}{2}})^2k = H(\theta,\theta')(\tau n^{-\frac{1}{2}})^2,$$

$$\min_{H(\theta,\theta')\geq 1} \frac{\|\Omega'(\theta) - \Omega'(\theta')\|_2^2}{H(\theta,\theta')} \geq (\tau n^{-\frac{1}{2}})^2.$$

Plugging the results of (2.22) and (2.23) into Lemma 6, we have

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_3} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \geq \frac{\tau^2}{16} n^{-1}k = \frac{\tau^2}{16} n^{-1} \min\{n^{\frac{1}{2\alpha+1}}, \frac{p}{2}\}.$$

APPENDIX G

PROOF FOR THEOREM 7

This proof is similar to that of Theorem 3. To prove Theorem 7, for any bandwidth k, according to (2.9), we have

$$\|\tilde{\Omega}_k^{\text{op}} - \Omega\|_{\text{op}}^2 \leq 8\|\tilde{\Omega}_k^* - \Omega_k^*\|_{\text{op}}^2 + 8\|\Omega_k^* - \Omega\|_{\text{op}}^2.$$

Lemma 2 indicates that

$$\|\Omega_k^* - \Omega\|_{\text{op}}^2 \leq Ck^{-2\alpha+1}.$$

Equations (2.10) - (2.12) together show that

$$\|\tilde{\Omega}_k^* - \Omega_k^*\|_{\text{op}}^2 \leq 16\eta^2 \max_{m\in[p]} \big(\|\mathbf{C}_m^{6k}(\frac{1}{n}\mathbf{Z}\mathbf{Z}^T) - \mathbf{C}_m^{6k}(\Omega^{-1})\|_{\text{op}} + \|\mathbf{C}_{k+1}^k\big((\mathbf{C}_m^{3k}(\Omega^{-1})^{-1})\big) - \mathbf{C}_m^k(\Omega)\|_{\text{op}}^2\big).$$

Note that Lemma 4 indicates

$$\|\mathbf{C}_{k+1}^k\big((\mathbf{C}_m^{3k}(\Omega^{-1})^{-1})\big) - \mathbf{C}_m^k(\Omega)\|_{\text{op}}^2 \leq Ck^{-2\alpha+1}.$$

In addition, the probability version of Lemma 3 (its proof can be found in Lemma 3 of [Cai et al., 2010]) indicates that for any $C_1 > 0$, one can find a sufficiently large $C > 0$ irrelevant of α such that

$$\mathbb{P}\big(\max_{m\in[p]} \|(\mathbf{C}_m^{6k}(\frac{1}{n}\mathbf{Z}\mathbf{Z}^T) - (\mathbf{C}_m^{6k}(\Omega^{-1})\|_{\text{op}}^2 \leq C\frac{\log p + k}{n}\big) \geq 1 - \exp\big(-C_1(\log p + k)\big).$$

Combining the above arguments, we derive the desired result in Lemma 15.

APPENDIX H

PROOFS FOR THEOREM 8

In this section, we prove Lemma 16, Lemma 17 to establish Theorem 8.

H.1 PROOF OF LEMMA 16

It is easy to show $\mathrm{diag}(\Omega^{-1}) = I$. Now we need to verify $\Omega \in \mathcal{P}_\alpha(\eta^2, M\eta)$. Let S denote $(\mathrm{diag}(\Omega'^{-1}))^{\frac{1}{2}}$. The Cholesky decomposition of Ω is:

$$\begin{aligned}\Omega &= S\Omega' S\\ &= S(I - A')^T D'^{-1}(I - A')S\\ &= (S - A'S)^T D'^{-1}(S - A'S)\\ &= (I - S^{-1}A'S)^T SD'^{-1}S(I - S^{-1}A'S).\end{aligned}$$

According to Lemma 24, $\eta^{-\frac{1}{2}} \le \lambda_{\min}(S) \le \lambda_{\max}(S) \le \eta^{\frac{1}{2}}$ and $\eta^{-1} \le \lambda_{\min}(\Omega') \le \lambda_{\max}(\Omega') \le \eta$. Therefore, we obtain $\eta^{-2} \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) \le \eta^2$. Let $A \equiv [a_{ij}]_{p\times p} = S^{-1}A'S$. The desired result $\Omega \in \mathcal{P}_\alpha(\eta^2, M\eta)$ then immediately follows from that $\max_i \sum_{j<i-k} |a_{ij}| < M\eta k^{-\alpha}$ for $k \in [i-1]$.

Since $\mathrm{diag}(\Omega'^{-1})_i^{\frac{1}{2}} > 0$, f_i is a strictly increasing function. Therefore, $\{\Omega, \{f_i\}\} \in \mathcal{P}'_\alpha(\eta^2, M\eta)$.

H.2 PROOF OF LEMMA 17

Set $I_k + A_k^*(\theta)^T A_k^*(\theta)$ as $W(\theta)$. One can check

$$\Omega(\theta) = \begin{bmatrix} W(\theta) & -A_k^*(\theta)^T \mathrm{diag}(W(\theta))^{\frac{1}{2}} & 0_{k\times(p-2k)} \\ -\mathrm{diag}(W(\theta))^{\frac{1}{2}} A_k^*(\theta) & \mathrm{diag}(W(\theta)) & 0_{k\times(p-2k)} \\ 0_{(p-2k)\times k} & 0_{(p-2k)\times k} & I_{p-2k} \end{bmatrix}.$$

Then,

$$\begin{aligned} \|\Omega(\theta') - \Omega(\theta)\|_{\mathrm{op}}^2 &\geq \|\mathrm{diag}(W(\theta))^{\frac{1}{2}} A_k^*(\theta) - \mathrm{diag}(W(\theta'))^{\frac{1}{2}} A_k^*(\theta')\|_{\mathrm{op}}^2 \\ &= H(\theta, \theta')(1 + (\tau n^{-\frac{1}{2}})^2)(\tau n^{-\frac{1}{2}})^2, \end{aligned}$$

which further implies

$$\min_{H(\theta,\theta')\geq 1} \frac{\|\Omega(\theta) - \Omega(\theta')\|_2^2}{H(\theta, \theta')} \geq (\tau n^{-\frac{1}{2}})^2.$$

APPENDIX I

PROOF FOR THEOREM 9

Let the linear projection of X_i onto the linear span of $\mathbf{X}_{i-k:i-1}$ in population be $\hat{X}_i^{\langle k\rangle}$. With the corresponding coefficients padded with $i-k-1$ zeros in the front, we can rewrite it as $\hat{X}_i^{\langle k\rangle} = \mathbf{X}_{1,i-1}^T \mathbf{a}_i^{\langle k\rangle}$, where $\boldsymbol{a}_i^{\langle k\rangle} \in \mathbb{R}^{i-1}$ with its first $i-k-1$ coordinates being zeros. In addition, we set $d_i^{\langle k\rangle} = \mathsf{Var}(X_i - \hat{X}_i^{\langle k\rangle})$. Note that $\mathbf{a}_i = \mathbf{a}_i^{\langle i-1\rangle}$, $d_i = d_i^{\langle i-1\rangle}$. Let $\hat{\mathbf{a}}_i^{\langle k_1\rangle}$ and $\hat{d}_i^{\langle k_1\rangle}$ be the empirical coefficient and residual of the regression $X_i \sim \mathbf{X}_{i-k_1:i-1}$, where $k_1 = \lceil \frac{n}{c} \rceil$ with some sufficiently large $c > 1$. The coefficients with threshold $\hat{\mathbf{a}}_i^{\langle k_1\rangle *}$ is defined as (3.2). According to (3.2) and (3.3), we know $\hat{\mathbf{a}}_i^* = \hat{\mathbf{a}}_i^{\langle k_1\rangle *}$, $\hat{d}_i = \hat{d}_i^{\langle k_1\rangle}$.

First, we prove that $\mathbb{E}|\hat{d}_i - d_i|^2 \le Cn^{-\frac{2\alpha+1}{2\alpha+2}}$. To this end, we decompose it as follow,

$$\begin{aligned}\mathbb{E}|\hat{d}_i - d_i|^2 &= \mathbb{E}|\hat{d}_i^{\langle k_1\rangle} - d_i^{\langle i-1\rangle}|^2 \\ &\le 2\mathbb{E}|\hat{d}_i^{\langle k_1\rangle} - d_i^{\langle k_1\rangle}|^2 + 2|d_i^{\langle k_1\rangle} - d_i^{\langle i-1\rangle}|^2.\end{aligned}$$

According to Lemma 25 we know that $|d_i^{\langle k_1\rangle} - d_i^{\langle i-1\rangle}|^2 \le Cn^{-2\alpha} \le Cn^{-\frac{2\alpha+1}{2\alpha+2}}$, noting that $k_1 = \lceil \frac{n}{c} \rceil$ and $\alpha > \frac{1}{2}$. Besides, the regression theory implies that, $(n-k_1)\hat{d}_i^{\langle k_1\rangle}/d_i^{\langle k_1\rangle} \sim \chi^2(n-k_1)$. So we have

$$\mathbb{P}(|\hat{d}_i^{\langle k_1\rangle}/d_i^{\langle k_1\rangle} - 1| > t) \le 2\exp(-(n-k_1)t^2/8), \quad t \in (0,1).$$

Then one can check that

$$\mathbb{E}|\hat{d}_i^{\langle k_1\rangle} - d_i^{\langle k_1\rangle}|^2 \le Cn^{-\frac{2\alpha+1}{2\alpha+2}}.$$

Therefore, we have shown that $\mathbb{E}|\hat{d}_i^{\langle k_1\rangle} - d_i|^2 \le Cn^{-\frac{2\alpha+1}{2\alpha+2}}$ by combining the above two equations together.

Then we turn to prove $\mathbb{E}\|\hat{\mathbf{a}}_i^* - \mathbf{a}_i\|_2^2 \leq Cn^{-\frac{2\alpha+1}{2\alpha+2}}$.

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{a}}_i^* - \mathbf{a}_i\|_2^2 =& \mathbb{E}\|\hat{\mathbf{a}}_i^{\langle k_1\rangle *} - \mathbf{a}_i^{\langle i-1\rangle}\|_2^2 \\ \leq& 2\mathbb{E}\|\hat{\mathbf{a}}_i^{\langle k_1\rangle *} - \mathbf{a}_i^{\langle k_1\rangle}\|_2^2 + 2\|\mathbf{a}_i^{\langle i-1\rangle} - \mathbf{a}_i^{\langle k_1\rangle}\|_2^2.\end{aligned}$$

According to Lemma 25, we know that $\|\mathbf{a}_i^{\langle i-1\rangle} - \mathbf{a}_i^{\langle k_1\rangle}\|_2^2 \leq Cn^{-2\alpha} \leq Cn^{-\frac{2\alpha+1}{2\alpha+2}}$, since $\alpha > \frac{1}{2}$. It is sufficient to prove that $\mathbb{E}\|\hat{\mathbf{a}}_i^{\langle k_1\rangle *} - \mathbf{a}_i^{\langle k_1\rangle}\|_2^2 \leq Cn^{-\frac{2\alpha+1}{2\alpha+2}}$.

We focus on the regression coefficients $\hat{\mathbf{a}}_i^{\langle k_1\rangle}$ first. The following analysis is conditioned on $\mathbf{Z}_{i-k_1:i-1}$. Note that with probability one that $(\mathbf{Z}_{i-k_1:i-1}^T\mathbf{Z}_{i-k_1:i-1})^{-1}$ exists since $k_1 = \lceil n/c\rceil$. For any fixed i, we have

$$\hat{\mathbf{a}}_i^{\langle k_1\rangle}|\mathbf{Z}_{i-k_1:i-1} \sim N(\mathbf{a}_i^{\langle k_1\rangle}, (\mathbf{Z}_{i-k_1:i-1}^T\mathbf{Z}_{i-k_1:i-1})^{-1}\mathsf{Var}(X_i|\mathbf{Z}_{i-k_1:i-1})).$$

For each coordinate in $\hat{\mathbf{a}}_i^{\langle k_1\rangle} = (\mathbf{0}, \hat{a}_{i(i-k_1)}^{\langle k_1\rangle} \ldots, \hat{a}_{i(i-1)}^{\langle k_1\rangle})^T$, $\hat{a}_{ij}^{\langle k_1\rangle}|\mathbf{Z}_{i-k_1:i-1}$ with $j \in i-k_1 : i-1$ can be represented as $a_{ij}^{\langle k_1\rangle} + \sigma_j$, where σ_j follows the normal distribution with variance which can be bounded as follows since $\Omega \in \mathcal{P}_\alpha(\eta, M)$,

$$\begin{aligned}\mathsf{Var}(\sigma_j) &\leq \|(\mathbf{Z}_{i-k_1:i-1}^T\mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\mathrm{op}}\mathsf{Var}(X_i|\mathbf{Z}_{i-k_1:i-1}) \\ &\leq \eta\|(\mathbf{Z}_{i-k_1:i-1}^T\mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\mathrm{op}}.\end{aligned} \tag{I.1}$$

Next, we can show,

$$|\hat{a}_{ij}^{\langle k_1\rangle *} - a_{ij}^{\langle k_1\rangle}| \leq \min\{|a_{ij}^{\langle k_1\rangle}|, \frac{3}{2}\lambda_j\} + |3\sigma_j|\mathbf{1}\left(|\sigma_j| > \frac{1}{2}\lambda_j\right). \tag{I.2}$$

To see this, one can check that

$$|\hat{a}_{ij}^{\langle k_1\rangle *} - a_{ij}^{\langle k_1\rangle}| = \max\{|\sigma_j|\mathbf{1}\left(|a_{ij}^{\langle k_1\rangle} + \sigma_j| > \lambda_j\right), |a_{ij}^{\langle k_1\rangle}|\mathbf{1}\left(|a_{ij}^{\langle k_1\rangle} + \sigma_j| \leq \lambda_j\right)\}.$$

For the first term, we have

$$\begin{aligned}&|\sigma_j|\mathbf{1}\left(|a_{ij}^{\langle k_1\rangle} + \sigma_j| > \lambda_j\right) \\ \leq& |\sigma_j|\mathbf{1}\left(|\sigma_j| > \lambda_j/2\right) + |\sigma_j|\mathbf{1}\left(|\sigma_j| \leq \lambda_j/2 \cap |a_{ij}^{\langle k_1\rangle}| > \lambda_j/2\right) \\ \leq& |\sigma_j|\mathbf{1}\left(|\sigma_j| > \lambda_j/2\right) + \min\{|a_{ij}^{\langle k_1\rangle}|, \lambda_j/2\}.\end{aligned}$$

Similarly, for the second term, we have

$$
\begin{aligned}
&|a_{ij}^{\langle k_1\rangle}|\mathbf{1}\left(|a_{ij}^{\langle k_1\rangle}+\sigma_j|\le\lambda_j\right)\\
\le&|a_{ij}^{\langle k_1\rangle}|\mathbf{1}\left(|a_{ij}^{\langle k_1\rangle}+\sigma_j|\le\lambda_j\cap|\sigma_j|>\lambda_j/2\right)\\
&+|a_{ij}^{\langle k_1\rangle}|\mathbf{1}\left(|a_{ij}^{\langle k_1\rangle}+\sigma_j|\le\lambda_j\cap|\sigma_j|\le\lambda_j/2\right)\\
\le&|3\sigma_j|\mathbf{1}\left(|\sigma_j|>\lambda_j/2\right)+\min\{|a_{ij}^{\langle k_1\rangle}|,3\lambda_j/2\}.
\end{aligned}
$$

We finish the proof of Equation (I.2).

Equation (I.2) further implies

$$
\begin{aligned}
&\mathbb{E}(\|\hat{\mathbf{a}}_i^{\langle k_1\rangle*}-\mathbf{a}_i^{\langle k_1\rangle}\|_2^2|\mathbf{Z}_{i-k_1:i-1})\\
\le&2\sum_{j=i-k_1}^{i-1}\min\{|a_{ij}^{\langle k_1\rangle}|,3/2\lambda_j\}^2+2\sum_{j=i-k_1}^{i-1}\mathbb{E}\big((3\sigma_j)^2\mathbf{1}\left(|\sigma_j|>\lambda_j/2\right)|\mathbf{Z}_{i-k_1:i-1}\big)\\
\le&4\sum_{j=i-k_1}^{i-1}\min\{|a_{ij}^{\langle i-1\rangle}|,3/2\lambda_j\}^2+18\sum_{j=i-k_1}^{i-1}\mathbb{E}(\sigma_j^2\mathbf{1}\left(|\sigma_j|>\lambda_j/2\right)|\mathbf{Z}_{i-k_1:i-1})\\
&+4\|\mathbf{a}_i^{\langle k_1\rangle}-\mathbf{a}_i^{\langle i-1\rangle}\|_2^2\\
\le&6\sum_{j=i-k_1}^{i-1}|a_{ij}^{\langle i-1\rangle}|\lambda_j+18\sum_{j=i-k_1}^{i-1}\mathbb{E}(\sigma_j^2\mathbf{1}\left(|\sigma_j|>\lambda_j/2\right)|\mathbf{Z}_{i-k_1:i-1})+Cn^{-\frac{2\alpha+1}{2\alpha+2}}.
\end{aligned}
\tag{I.3}
$$

Set J_0 as $\log_2^{k_0}$, J_1 as $\log_2^{k_1}$. Due to that $\Omega\in\mathcal{P}_\alpha(\eta,M)$ and (I.1), we can show

$$
\begin{aligned}
&\sum_{j=i-k_1}^{i-1}|a_{ij}^{\langle i-1\rangle}|\lambda_j\\
\le&\sum_{k=J_0}^{J_1+1}\sqrt{(k-J_0)R}\times M(2^{-(k-J_0)\alpha}-2^{-(k+1-J_0)\alpha})2^{-J_0\alpha}\\
\le&n^{-1/2}2^{-J_0\alpha}M\sqrt{8\eta\|n(\mathbf{Z}_{i-k_1:i-1}^T\mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\text{op}}}\Big(\sum_{k=0}^{J_1-J_0+1}2^{-k\alpha}\sqrt{k}\Big)\\
\le&Cn^{-\frac{2\alpha+1}{2\alpha+2}}\sqrt{\|n(\mathbf{Z}_{i-k_1:i-1}^T\mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\text{op}}}\\
\le&C(\|n(\mathbf{Z}_{i-k_1:i-1}^T\mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\text{op}}+1)n^{-\frac{2\alpha+1}{2\alpha+2}},
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_{j=i-k_1}^{i-1} \mathbb{E}(\sigma_j^2 \mathbf{1}\left(|\sigma_j| > \lambda_j/2\right) |\mathbf{Z}_{i-k_1:i-1}) \\
\leq &\sum_{j=i-k_1}^{i-1} \mathbb{E}(\eta \|(\mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\text{op}} \mathbf{1}\left(|\sigma_j| > \lambda_j/2\right) |\mathbf{Z}_{i-k_1:i-1}) \\
\leq &\eta \|(\mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\text{op}} \sum_{j=i-k_1}^{i-1} \mathbb{P}(|\sigma_j| > \lambda_j/2 |\mathbf{Z}_{i-k_1:i-1}) \\
\leq &n^{-1} C \|n(\mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\text{op}} \sum_{j=i-k_1}^{i-1} \exp(-\frac{\lambda_j^2}{8\eta \|(\mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\text{op}}}) \\
\leq &n^{-1} C \|n(\mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\text{op}} \sum_{k=J_0}^{J_1+1} \exp(-(k - J_0))(2^k) \\
\leq &C \|n(\mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\text{op}} n^{-\frac{2\alpha+1}{2\alpha+2}}.
\end{aligned}
$$

Plugging the above two equations in (I.3), we have

$$
\begin{aligned}
&\mathbb{E}(\|\hat{\mathbf{a}}_i^{\langle k_1 \rangle *} - \mathbf{a}_i^{\langle k_1 \rangle}\|_2^2) \\
= &\mathbb{E}(\mathbb{E}(\|\hat{\mathbf{a}}_i^{\langle k_1 \rangle *} - \mathbf{a}_i^{\langle k_1 \rangle}\|_2^2 |\mathbf{Z}_{i-k_1:i-1})) \\
\leq &C(\mathbb{E}\|n(\mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\text{op}} + 1) n^{-\frac{2\alpha+1}{2\alpha+2}} \\
&+ C\mathbb{E}\|n(\mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\text{op}} n^{-\frac{2\alpha+1}{2\alpha+2}} + C n^{-\frac{2\alpha+1}{2\alpha+2}} \\
\leq &C n^{-\frac{2\alpha+1}{2\alpha+2}}.
\end{aligned}
$$

The last inequality holds since $\mathbb{E}\|n(\mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\text{op}}$ can be bounded by constant, noting that $k_1 = \lceil n/c \rceil$ with some sufficiently large $c > 1$. Indeed, one can follow the strategy in the proof of Theorem 5 in Cai et al. [2010], together with the concentration inequality of the sample covariance matrix under the operator norm (e.g., from Theorem 5.39 in Vershynin [2010]). It follows that

$$
\mathbb{E}(\|\hat{\mathbf{a}}_i^* - \mathbf{a}_i\|_2^2) \leq C n^{-\frac{2\alpha+1}{2\alpha+2}}.
$$

APPENDIX J

PROOFS FOR THEOREM 10

In this section, we prove Lemma 20, Lemma 21 to establish Theorem 10.

J.1 PROOF OF LEMMA 20

Let $C_s(\theta(s)) \equiv [c(s)_{ij}]_{k\times k}$, it is easy to check $\sum_{i-j>k} |c_{ij}| \leq Mk^{-\alpha-1}$. One can check $\Omega(\theta)$ in $\mathcal{P}_4$ has the specific form of

$$\Omega(\theta) = \begin{bmatrix} E_1(\theta(1)) & 0_{2k} & \dots & 0_{2k} \\ 0_{2k} & E_2(\theta(2)) & \dots & 0_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{2k} & 0_{2k} & \dots & E_{\lceil \frac{p}{2k} \rceil}(\theta(\lceil \frac{p}{2k} \rceil)) \end{bmatrix},$$

where,

$$E_s = \begin{bmatrix} I_k + C_s(\theta(s))^T C_s(\theta(s)) & -C_s(\theta(s))^T \\ -C_s(\theta(s)) & I_k \end{bmatrix}.$$

Let $\mathbf{1}$ denote the vector with all 1's in Θ. Then one can check

$$\begin{aligned} \lambda_{\max}(\Omega(\theta)) =& \big(\lambda_{\max}(I - C(\theta))\big)^2 \leq \big(\lambda_{\max}(I + C(\theta))\big)^2 \leq (1 + \lambda_{\max}(C_s(\mathbf{1})))^2 \\ =& (1 + kn^{-\frac{1}{2}}\tau)^2 \leq \eta. \end{aligned}$$

The second inequality above is due to fact that the entries of $C(\theta)$ are all non-negative and Lemma 27. Recall $\Sigma(\theta) = (I + C(\theta))(I + C(\theta))^T$. Therefore, we have

$$\begin{aligned}\lambda_{\min}(\Omega(\theta)) =& \big(\lambda_{\max}(\Sigma(\theta))\big)^{-1} = \big(\lambda_{\max}(I + C(\theta))\big)^{-2} \geq (1 + \lambda_{\max}(C_s(\mathbf{1})))^{-2} \\ =& (1 + kn^{-\frac{1}{2}}\tau)^{-2} \geq \eta^{-1}.\end{aligned}$$

The eigenvalues of $\Omega(\theta)$ are in the interval $[\eta^{-1}, \eta]$. So $\mathcal{P}_4 \in \mathcal{P}_\alpha(\eta, M)$.

J.2 PROOF OF LEMMA 21

For (3.11), the proof is almost the same as that of Lemma 7. Since Lemma 21 is about the distributions in the subset $\mathcal{P}_4$, we need to recheck the inequalities (D.1) and (D.2). Assume that $\Sigma(\theta) = \{\Omega(\theta)^{-1} : \Omega(\theta) \in \mathcal{P}_4\}$ and $D = \Sigma(\theta') - \Sigma(\theta)$, note that $H(\theta, \theta') = 1$, in this case, one can verify that we only need to replace A_k^* by C_s. Specifically, C_s corresponds to the one where θ and θ' are different. Then we still have

$$\begin{aligned}\|D\|_{\text{op}} &\leq 2(1 + \|C_s(\mathbf{1})\|_{\text{op}})\|C_s(d)\|_{\text{op}} \leq 4\tau n^{-\frac{1}{2}}, \\ \|D\|_{\text{F}} &\leq 2(1 + \|C_s(\mathbf{1})\|_{\text{F}})\|C_s(d)\|_{\text{F}} \leq 4\tau n^{-\frac{1}{2}}.\end{aligned}$$

So (3.11) holds in this case. As for (3.12), similarly,

$$\begin{aligned}\|\Omega(\theta') - \Omega(\theta)\|_{\text{F}}^2 \geq \sum_s \|C_s(\theta'(s) - \theta(s))\|_{\text{F}}^2 = H(\theta, \theta')(\tau n^{-\frac{1}{2}})^2, \\ \min_{H(\theta,\theta')\geq 1} \frac{\|\Omega'(\theta) - \Omega'(\theta')\|_2^2}{H(\theta, \theta')} \geq \tau^2 n^{-1}.\end{aligned}$$

Proofs of Lemma 16 and Lemma 17 in the analysis of Theorem 8 In this section, we prove Lemma 16 and Lemma 17 to establish Theorem 8.

APPENDIX K

R CODE FOR SECTION 4.1

```
############## setting list #####################
eta=10000
M=2
d=1
p.list=c(500, 1000, 2000)
alpha.list=c(0.5, 1, 1.5, 2)
n.list=c(500, 1000, 2000, 4000)
total=100

library(MASS)
############## generating data ############################
omega.f<-function(p, M, alpha, d){
  A=matrix(0,p,p)
  for (i in 2:p){
      for (j in 1:(i-1)){
          A[i,1]= M * (abs(i-j))^{- alpha-1}
       }
  }
```

```
  A[row(A)==col(A)]=1
  D=diag(rep(d^{-1},p))
  omega=t(A) %*% D %*% A
  return (omega)}

############# method in the thesis ##########################
index <- function (n, p){
  if (n < 1) {return (1)}
  if (n > p) {return (p)}
  return (n)
}
restrict <- function (m, eta){
  for (i in 1: length(m))
  {
    if (m[i] < eta^-1) m[i]=eta^-1
    if (m[i] > eta) m[i]=eta
  }
  return(m)
}
proj <- function (m, eta){
  s <- eigen(m)
  D <- diag(restrict(s$values, eta))
  return (s$vectors %*% D %*% solve(s$vectors))
}


tapering.k.est <-function(k, data, eta){
  if (k < 2) { k = 1 }
  p=dim(data)[2]
  n=dim(data)[1]
```

```
  est=matrix(0,p,p)
  for (i in (2-k):p){
    subindex=c(index((i-k),p):index((i+2*k-1),p))
    subset=data[,subindex]
    hat.cov= 1/n * t(subset) %*% subset
    tilde.cov=proj(hat.cov, eta)
    #tilde.cov = hat.cov
    temp1 = -index(i-k,p)+index(i,p)+1
    temp2 = index(i+k-1,p)-index(i-k,p)+1
    center=c(temp1 : temp2)
    tilde.pre=matrix(0,p,p)
    row = c(index(i,p):(index(i,p)+length(center)-1))
    col = c(index(i,p):(index(i,p)+length(center)-1))
    tilde.pre[row , col]=solve(tilde.cov)[center,center]
    est=est+tilde.pre
  }
  return(est)
}

est.z<-function(k, data, eta){
  part1 = tapering.k.est(k, data, eta)
  part2 = tapering.k.est(floor(k/2), data, eta)
  estimator=1/(k-floor(k/2))*(part1 - part2)
  return(estimator)
}
################ method in Bickel and Levina ############
index.f<-function(index,p){
  if (index > p) return(p)
  if (index < 1) return(1)
  return (index)
```

```
}

mse.f<-function(lm,n,k){
  return (sum(lm$residuals^2)/(n-k))
}

est.b <- function(k, data){
  n=dim(data)[1]
  p=dim(data)[2]
  A <- matrix(0,p,p)
  A[row(A)==col(A)] <- 1
  D=rep(0,p)
  D[1]=var(data[,1])/(n-1)*n
  for (i in 2:p){
    y <- data[,i]
    x <- data[,index.f(i-k, p):index.f(i-1, p)]
    lm <- lm(y ~ x - 1)
    D[i] <- mse.f(lm, n, (index.f(i-1, p)-index.f(i-k, p)))
    coeff <- lm$coefficients
    temp <- matrix(0, p, p)
    temp[i,index.f(i-k, p):index.f(i-1, p)]=-coeff
    A <- A + temp
  }
  D=diag(D^(-1))
  return (t(A) %*% D %*% A)}
########## bandwidth calculator ##########################
bd.b.f<-function (n,alpha,p)
 {return (floor((n/log(p))^{1/(2*alpha+2)}))}
bd.e.f<-function (n,alpha)
{return (floor(n^{1/(2*alpha+1)}))}
```

```
########## lets play ########################################
dist.f <- function(matrix)
{ return (max(abs(eigen(matrix)$values)))}

for (p in p.list){
  for (alpha in alpha.list){
    for (n in n.list){
      omega=omega.f(p, M, alpha, d)
      sigma=solve(omega)
      bd.b=bd.b.f(n,alpha,p)
      bd.e=bd.e.f(n,alpha)
      omega.eigen=dist.f(omega)

      bd.b.m.b1=rep(0,total)
      bd.e.m.z1=rep(0,total)

      for (k in 1:total){
        data=mvrnorm(n,rep(0,p),sigma)
        bd.b.m.b1[k]=dist.f(est.b(bd.b, data)-omega)
        bd.e.m.z1[k]=dist.f(est.z(bd.e, data, eta)-omega)
      }
      result=data.frame(bd.e.m.z1, bd.b.m.b1)
      write.csv(result, paste(p,n,alpha,"entry.csv"))
    }
  }
}
```

APPENDIX L

R CODE FOR SECTION 4.2

```
############### setting list ######################
eta=10000
M=2
d=1
p.list=c(500, 1000, 2000)
alpha.list=c(1, 1.5, 2)
n.list=c(500, 1000, 2000, 4000)
total=100

library(MASS)
############### generating data ##############################
omega.f<-function(p, M, alpha, d){
  A=matrix(0,p,p)
  for (i in 2:p){
   A[i,1]= M * (abs(i-1))^{- alpha}
   for (j in 2:(i-1)){
    A[i,j]= M * (abs(i-j))^{- alpha} - M * (abs(i-j)+1)^{- alpha}
    }}
```

```
  A[row(A)==col(A)]=1
  D=diag(rep(d^{-1},p))
  omega=t(A) %*% D %*% A
  return (omega)}

############## method in this thesis ##########################
index <- function (n, p){
  if (n < 1) {return (1)}
  if (n > p) {return (p)}
  return (n)
}
restrict <- function (m, eta){
  for (i in 1: length(m))
  {
    if (m[i] < eta^-1) m[i]=eta^-1
    if (m[i] > eta) m[i]=eta
  }
  return(m)
}
proj <- function (m, eta){
  s <- eigen(m)
  D <- diag(restrict(s$values, eta))
  return (s$vectors %*% D %*% solve(s$vectors))
}


tapering.k.est <-function(k, data, eta){
  if (k < 2) { k = 1 }
  p=dim(data)[2]
  n=dim(data)[1]
```

```
  est=matrix(0,p,p)
  for (i in (2-k):p){
    subindex=c(index((i-k),p):index((i+2*k-1),p))
    subset=data[,subindex]
    hat.cov= 1/n * t(subset) %*% subset
    tilde.cov=proj(hat.cov, eta)
    #tilde.cov = hat.cov
    temp1 = -index(i-k,p)+index(i,p)+1
    temp2 = index(i+k-1,p)-index(i-k,p)+1
    center=c(temp1 : temp2)
    tilde.pre=matrix(0,p,p)
    row = c(index(i,p):(index(i,p)+length(center)-1))
    col = c(index(i,p):(index(i,p)+length(center)-1))
    tilde.pre[row , col]=solve(tilde.cov)[center,center]
    est=est+tilde.pre
  }
  return(est)
}

est.z<-function(k, data, eta){
  part1 = tapering.k.est(k, data, eta)
  part2 = tapering.k.est(floor(k/2), data, eta)
  estimator=1/(k-floor(k/2))*(part1 - part2)
  return(estimator)
}
############ method in Bickel and Levina ####################
index.f<-function(index,p){
  if (index > p) return(p)
  if (index < 1) return(1)
  return (index)
```

```
}

mse.f<-function(lm,n,k){
  return (sum(lm$residuals^2)/(n-k))
}

est.b <- function(k, data){
  n=dim(data)[1]
  p=dim(data)[2]
  A <- matrix(0,p,p)
  A[row(A)==col(A)] <- 1
  D=rep(0,p)
  D[1]=var(data[,1])/(n-1)*n
  for (i in 2:p){
    y <- data[,i]
    x <- data[,index.f(i-k, p):index.f(i-1, p)]
    lm <- lm(y ~ x - 1)
    D[i] <- mse.f(lm, n, (index.f(i-1, p)-index.f(i-k, p)))
    coeff <- lm$coefficients
    temp <- matrix(0, p, p)
    temp[i,index.f(i-k, p):index.f(i-1, p)]=-coeff
    A <- A + temp
  }
  D=diag(D^(-1))
  return (t(A) %*% D %*% A)}

############### bandwidth calculator ##########################
bd.z.f<-function (n,alpha)
{return (floor(n^{1/(2*alpha)}))}
```

```
bd.b.f<-function (n,alpha,p)
 {return (floor((n/log(p))^{1/(2*alpha+2)}))}

bd.e.f<-function (n,alpha)
{return (floor(n^{1/(2*alpha+1)}))}

############## lets play ##############################
dist.f <- function(matrix)
{ return (max(abs(eigen(matrix)$values)))}

for (p in p.list){
  for (alpha in alpha.list){
    for (n in n.list){
      omega=omega.f(p, M, alpha, d)
      sigma=solve(omega)
      bd.z=bd.z.f(n,alpha)
      bd.b=bd.b.f(n,alpha,p)
      bd.e=bd.e.f(n,alpha)
      omega.eigen=dist.f(omega)

      bd.z.m.z1=rep(0,total)
      bd.b.m.b1=rep(0,total)
      bd.e.m.z1=rep(0,total)

      for (k in 1:total){
        data=mvrnorm(n,rep(0,p),sigma)
        bd.z.m.z1[k]=dist.f(est.z(bd.z, data, eta)-omega)
        bd.b.m.b1[k]=dist.f(est.b(bd.b, data)-omega)
        bd.e.m.z1[k]=dist.f(est.z(bd.e, data, eta)-omega)
      }
```

```
        result=data.frame(bd.z.m.z1, bd.e.m.z1, bd.b.m.b1)
        write.csv(result, paste(p,n,alpha,".csv"))
      }
    }
  }
```

BIBLIOGRAPHY

Patrice Assouad. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296(23):1021–1024, 1983.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre dAspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

Sayantan Banerjee and Subhashis Ghosal. Posterior convergence rates for estimating large precision matrices using graphical models. *Electronic Journal of Statistics*, 8(2):2111–2137, 2014.

Rina Foygel Barber and Mladen Kolar. Rocket: Robust confidence intervals via Kendall's tau for transelliptical graphical models. *arXiv preprint arXiv:1502.07641*, 2015.

Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008a.

Peter J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008b.

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Y Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

Jacob Bien, Florentina Bunea, and Luo Xiao. Convex banding of the covariance matrix. *Journal of the American Statistical Association*, 111(514):834–845, 2016.

T Tony Cai. Minimax and adaptive inference in nonparametric function estimation. *Statistical Science*, 27(1):31–50, 2012.

T Tony Cai and Harrison H Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420, 2012.

T Tony Cai, Cun-Hui Zhang, and Harrison H Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

T Tony Cai, Zhao Ren, and Harrison H Zhou. Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probability Theory and Related Fields*, 156(1-2):101–143, 2013.

T Tony Cai, Weidong Liu, and Harrison H Zhou. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2):455–488, 2016a.

T Tony Cai, Zhao Ren, and Harrison H Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59, 2016b.

Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011.

Tony Cai, Zongming Ma, and Yihong Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability theory and related fields*, 161(3-4):781–815, 2015.

I Csiszár. Information-type indices of the divergence of distributions. i, ii. *Magyar Tud. Akad. Mat. Fiz. Oszt. Közl. 17 (1967), 123-149; ibid*, 17:267–291, 1967.

Alexandre d'Aspremont, Onureena Banerjee, and Laurent El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1): 56–66, 2008.

Bernard Delyon and Anatoli Juditsky. On minimax wavelet estimators. *Applied and Computational Harmonic Analysis*, 3(3):215–228, 1996.

Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

Noureddine El Karoui. On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n tend to infinity. *arXiv preprint math/0309355*, 2003.

Noureddine El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, 36(6):2717–2756, 2008.

Jianqing Fan, Lingzhou Xue, and Hui Zou. Multitask quantile regression under the transnormal model. *Journal of the American Statistical Association*, 111(516):1726–1735, 2016.

Reinhard Furrer and Thomas Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255, 2007.

Thomas M Hamill, Jeffrey S Whitaker, and Chris Snyder. Distance-dependent filtering of background error covariance estimates in an ensemble kalman filter. *Monthly Weather Review*, 129(11):2776–2790, 2001.

Mark H Heyer and F Peter Schloerb. Application of principal component analysis to large-scale spectral line imaging studies of the interstellar medium. *The Astrophysical Journal*, 475(1):173, 1997.

Addison Hu and Sahand Negahban. Minimax estimation of bandable precision matrices. In *Advances in Neural Information Processing Systems*, pages 4893–4901, 2017.

Jianhua Z Huang, Naiping Liu, Mohsen Pourahmadi, and Linxu Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, pages 295–327, 2001.

Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486): 682–693, 2009.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254, 2009.

Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5): 603–621, 2003.

Kyoungjae Lee and Jaeyong Lee. Estimating large precision matrices via modified Cholesky decomposition. *arXiv preprint arXiv:1707.01143*, 2017.

O.V. Lepskii. Asymptotically minimax adaptive estimation. I: Upper bounds. Optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992.

Elizaveta Levina, Adam Rothman, and Ji Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, 2(1):245–263, 2008.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10 (Oct):2295–2328, 2009.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

Ritwik Mitra and Cun-Hui Zhang. Multivariate analysis of nonparametric estimates of large correlation matrices. *arXiv preprint arXiv:1403.6195*, 2014.

Nikhil Padmanabhan, Martin White, Harrison H Zhou, and Ross O'Connell. Estimating sparse precision matrices. *Monthly Notices of the Royal Astronomical Society*, 460(2): 1567–1576, 2016.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617, 2007.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43 (3):991–1026, 2015.

Adam J Rothman, Peter J Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Leonas Saulis and VA Statulevicius. *Limit theorems for large deviations*, volume 73. Springer Science & Business Media, 2012.

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

Charles Spearman. Spearmans rank correlation coefficient. *Amer J Psychol*, 15:72–101, 1904.

Tingni Sun and Cun-Hui Zhang. Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research*, 14(1):3385–3418, 2013.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Kenneth W Wachter. Probability plotting points for principal components. In *Ninth Interface Symposium Computer Science and Statistics*, pages 299–308. Prindle, Weber and Schmidt, Boston, 1976.

Kenneth W Wachter. The strong limits of random matrix spectra for sample matrices of independent elements. *The Annals of Probability*, 6(1):1–18, 1978.

Wei Biao Wu and Mohsen Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003.

Wei Biao Wu and Mohsen Pourahmadi. Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, pages 1755–1768, 2009.

Luo Xiao and Florentina Bunea. On the theoretic and practical merits of the banding estimator for large covariance matrices. *arXiv preprint arXiv:1402.0844*, 2014.

Bin Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, 423:435, 1997.

Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286, 2010.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.